An Integrated Model of Semantics and Control, Part 2:

Solving the Similarity Paradox Through Context Inference

Tyler Giallanza[*]

Timothy T. Rogers[‡]

Jonathan D. Cohen[*†]

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

[‡]*Department of Psychology*
University of Wisconsin, Madison

Keywords: category learning; cognitive control; coherent covariation; connectionism; context processing; generalization; induction; large language models; LLMs; neural networks; statistical learning.

Acknowledgements

The authors would like to thank Declan Campbell for helpful discussions that inspired the work reported in this article. The work reported in this article was supported by the following grants: NSF Graduate Fellowship (TG); NSF Award 21-517 (TTR); Vannevar Bush Fellowship sponsored by ONR (JDC).

All data and code to reproduce the simulations are available on GitHub.
The experiments in this article were not preregistered.

Abstract

Semantic similarity plays an ambiguous role in models of human cognition. On the one hand, it is often viewed as a foundational construct that shapes how we categorize, learn, and make inferences about objects and their properties. On the other hand, a host of behavioral evidence suggests that similarity is too rigid to explain the flexibility of inductive inference. We present the Integrated Semantics and Control — Context Inference (ISC-CI) model to resolve this tension, proposing that flexible inference emerges within a system that dynamically reshapes represented semantic similarities amongst stimuli depending upon the immediate context. The ISC-CI model builds on prior models of semantics and control that learn how to build and flexibly access semantic knowledge from observing the statistical relationships between objects, their properties, and the contexts in which these occur. Critically, it introduces a new mechanism that infers a suitable representation of context for both familiar and novel scenarios, without any direct labeling in the environment. The inferred context allows the system to selectively weight different dimensions within its representational space depending on the items being processed. Through simulations and experiments, we demonstrate that the ISC-CI model provides a coherent account of performance across inductive inference and semantic similarity tasks, including classic tasks that have long challenges theories of induction, offering a unified account of these cognitive processes that highlights the importance of context. We conclude by considering the implications of these findings for broader questions in cognitive science and artificial intelligence.

Introduction

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

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

Although similarity-only models explain a wide range of behavioral findings, a set of exceptions suggests that similarity alone is insufficient to fully explain human inductive inference. For example, induction of novel category labels is sensitive to the distribution of labeled examples (Xu & Tenenbaum, 2007): When shown a single green pepper with the label "Fep," most adults extend the label to other varieties of peppers; however, when shown three green peppers all labeled "Fep," adults limit the label extension only to other green peppers. This result is difficult to explain solely from the conceptual similarity between the items, which should be near-identical for one versus three green peppers. Furthermore, judgments of similarity themselves appear to violate axioms of the similarity-only approach: People produce asymmetric similarity ratings for pairs of items depending on the order they are listed, in ways that reflect typicality (e.g., generating higher ratings when asked how similar donkeys are to horses than when asked how similar horses are to donkeys), and multi-alternative similarity judgments can result in preference reversals (e.g., when asked which of Paris, Berlin, or York is most similar to London, most people choose York, suggesting that York is more similar to London than is Paris; but when given the options of Paris, *Liverpool*, or York, most people choose Paris, suggesting the opposite conclusion; Tversky & Gati, 1978).

Such findings are often interpreted as illustrating that human induction requires qualitatively different kinds of representations and processing mechanisms in addition to similarity-based generalization alone. For instance, Osherson's influential Similarity-Coverage Model combines a similarity score with a "coverage" term derived from discrete, mutually-exclusive, taxonomically-defined category representations (Osherson et al., 1990). Similarly, Xu and Tenenbaum's (2007) Bayesian inference model, designed to account for category label induction, relies on the similarity between objects, but it calculates these similarities using discrete categories situated within a taxonomic hierarchy. Tversky's feature contrast model, devised to account for asymmetric similarity judgments, calculates similarity, but it does so by representing the features of concepts as discrete sets that then receive different weightings in the final judgment depending on how the question is framed (Tversky, 1977; Tversky & Gati,

1978). Each of these models has been successful in accounting for the data they were meant to address, but it is not clear how these different models relate to one another, or to the many other empirical observations of human behavior for which similarity alone *does* seem to be a sufficient account. To our knowledge, no single approach has accounted for the full variety of relevant phenomena, and some phenomena, such as preference reversal in multi-alternative similarity judgments, have not been explained by any prior formal model.

In this article, we propose that the tension between similarity-only and similarity-plus approaches can be reconciled under an account of human inductive inference in which: a) inferences are always supported by similarity within a continuous, metric conceptual representational space; but b) the distance between representations within the space, and hence their relative similarities to one another, is context-sensitive and subject to dynamic cognitive control. Variants of this idea have abounded, including in Tversky's own thinking on the question (Tversky, 1977), but the structure of the context representations needed to modulate semantic relationships has never been made clear, nor is it clear how these might be acquired, how and under what circumstances they are deployed, or how participants can infer the appropriate context representations to use "on the fly" without special instruction in a given task setting — that is, how context relates to control. We show how a neural network model, based on statistical learning of both semantic and context representations, that is subject to control through both the learning and online construction and use of context representations, can address these questions and explain the range of phenomena listed above. We then report new experiments designed to test specific predictions of the model that contrast with other classic and contemporary models.

Our approach builds on foundational work on neural network models of semantics (Rogers & McClelland, 2004; Rumelhart & Todd 1993) and recent efforts to integrate such models with similarly cast models of cognitive control (Giallanza et al., 2024; Lambon Ralph et al., 2017). The basic premises of this work are that: a) semantic knowledge is acquired through statistical learning; b) it is shaped in use by the influence of context representations that reflect current

instructions and/or behavioral demands; and c) these context representations are themselves subject to the same mechanisms of statistical learning, spanning multiple levels of abstraction, and driven by behavioral affordances together with perceptual statistics.

Recent work has suggested how a model of such an integrated semantic and control system might acquire representations of both conceptual structure and of task contexts through learning, and how the two forms of representation might jointly contribute to the representational structure that ultimately drives overt, semantically informed and contextually-appropriate behaviors in a given task (Giallanza et al., 2024). However, that model, and previous ones on which it was based (e.g., Rogers & McClelland, 2004), all made the critical simplifying assumption that the "tasks" the system carries out are somehow labelled in the input; that is, the system is externally "instructed" about the kind of information it is to report.

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

Here, we develop a mechanistic hypothesis about how an integrated semantic/control system can acquire useful task/context representations without these being directly labeled by

the environment, and how such a system can then infer which context representations are most useful on the fly without such instruction. We begin with a brief review of the Integrated Semantics and Control (ISC; Giallanza et al., 2024) model on which the present work builds. We next explain how the ISC framework can be extended to learn and inter context representations on its own, without requiring direct labeling of task contexts in the input. We then describe a neural network model that extends the ISC framework by *inferring a useful contextual representation* in a given situation: the ISC Context-Inference model, or ISC-CI.

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

Integrating Semantics and Control through Contextual Inference

Background: Integrated Semantics and Cognition

Our proposal extends prior work that takes a statistical approach to understanding how semantic knowledge is acquired, structured, accessed, and used (Giallanza et al., 2024; Hinton, 1981; 1986; Rogers & McClelland, 2004; Rumelhart & Todd, 1993). This work proposes that semantic knowledge is *acquired* by observing patterns of co-occurrence across objects and their properties, as well as the appropriate actions to take in response to them, under

different scenarios and contexts throughout development. Learning these patterns results in knowledge about how items relate to one another and to associated actions, because item properties are not distributed at random, but tend to co-occur together within similar types of things or in similar settings (Rosch, 1975). For example, the properties *has a beak*, *has-feathers*, and *has-wings* tend to all be present together in items we label "bird," but not in other items. This tendency for many properties to co-occur together across concepts has been termed *coherent covariation*, and prior work has shown that learning such patterns can support inferences based on partial information (e.g., if a new item has a beak and feathers, it is probably a bird) and relational judgments between objects (e.g., if two different objects both have beaks and feathers, they are probably similar to one another overall; Rogers & McClelland 2004; 2005).

In many neural network models of semantics, knowledge about objects, their properties, and the coherent covariation among these is implicitly encoded within the weights of a network that learns to generate correct inferences about an item from a subset of its observed properties. Perception of an item's name or other properties gives rise to distributed patterns of activation over units that serve as learned internal representations; this activation then propagates forward to generate outputs representing the system's inferences about and/or overt responses to the item's unobserved properties. After learning to generate correct inferences for many items, the internal representations capture important elements of *semantic structure:* items that are semantically related give rise to similar internal patterns of activation so that the ensemble of units can be viewed as capturing, within a distributed and multidimensional representational space, information about the semantic similarity relations amongst concepts. Items possessing similar properties will be represented with similar patterns of activity; put differently, semantically similar items will be nearby in the representational space. Proximity in the representational space in turn supports both inductive generalization (properties known to be true of one item will tend to generalize to nearby items)

and relational judgments (people will judge items nearby in the space to be similar kinds of things).

While many neural network models of semantic representation share these characteristics (e.g. McClelland & Farah, 1991; McRae et al., 1997; Plaut & Shallice, 1993; Seidenberg & McClelland, 1989), the I*ntegrated Semantics and Control* (ISC; Giallanza et al., 2024) framework (and related prior approaches; e.g., Rogers & McClelland, 2004) additionally suggests that the semantic space governing inference can be selectively *warped* or *controlled* by mechanisms responsive to current goals or task demands (a critical function of cognitive control; Miller & Cohen, 2001), so as to ensure that inferences and/or overt behaviors are suited to the corresponding task or context. This proposal was motivated by the long-standing observation that people discern and deploy different similarity relations amongst a given set of concepts depending on the current task (Saffran et al., 1996). For example, consider that ravens, robins, skunks, and foxes vary in both taxonomy (birds vs mammals) and color (black vs red/brown). The taxonomic relations guide inference for many kinds of properties: The observed shape, parts, behaviors, diet, category label, or genes of a raven should generalize more strongly to a robin than to a skunk, for instance. In some contexts, however, color is more important than taxonomy: when designing the composition of a painting, the raven and skunk might be represented as similar to one another and distinct from the robin or fox by virtue of their shared color. In the ISC framework, the controlled semantic system resolves the tension between these two representational structures (taxonomic vs color-based) by allowing a representation of the current task (which may be external, such as instructions from an experimenter to classify objects by color, or internal, such as a goal to finish a painting) to reshape the similarity relations expressed within the semantic representation space so as to emphasize similarities encoded by task-relevant properties and de-emphasize similarities encoded across features not relevant to the current task. Thus for designing the composition of a painting, the task representation might warp the expressed semantic representations so that similarity in color is more strongly expressed than is taxonomic similarity.

Giallanza et al. (2024) recently demonstrated how these principles are expressed in a simple feed-forward neural network model (Figure 1A) implemented within the ISC framework, that scales historical work on both semantics (e.g., Rogers & McClelland, 2004; Rumelhart & Todd, 1993) and control (e.g., Cohen et al., 1990) to a large, naturalistic dataset. The ISC model learns about objects and their properties in various contexts, producing representations that encode cross-context semantic structure in a *context independent* layer, information about the current context/task in a *context* layer, and context-relevant semantic structure in a *context dependent* layer. The model then uses this information to output properties that are true of a given object and relevant to the given task.

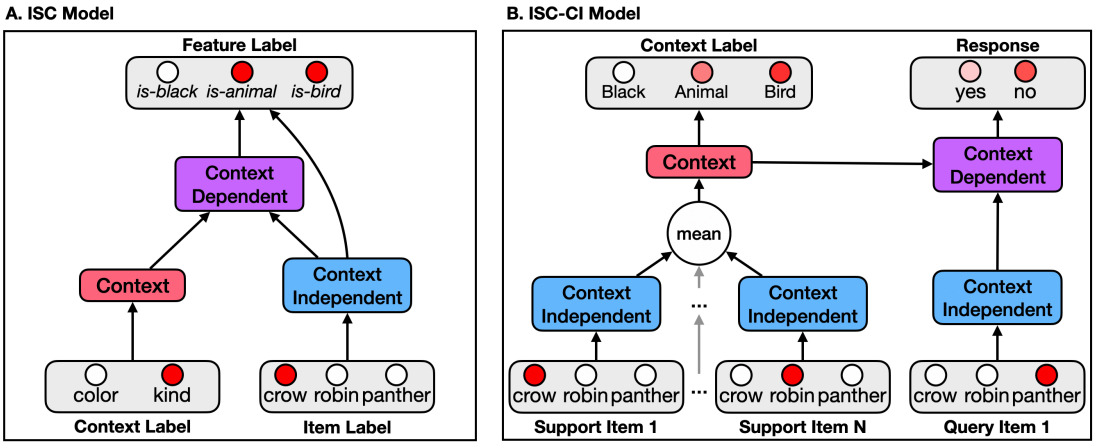

Figure 1. The ISC and ISC-CI models. A. The ISC model takes pre-specified item and task/context representations as input and learns to activate the features that are both true of the item and appropriate for the current context. It learns to represent cross-context item properties (context independent layer), information about the current context (context layer), and information about the context-relevant properties of the current item (context dependent layer). B. Building on the ISC model, the ISC-CI model takes a set of support items and a set of query items (only one of which is shown in the figure) as input. It learns to use the mean of the items in the support set to predict both what context it is in and, for each item in the query set, whether it can be expected to occur in that context.

As shown in Figure 1A, the model receives as input the current object (e.g., crow, raven, etc.) and task (e.g., *report name*, *report behavior*, etc.), represented as one-hot encodings. It produces as output a binary vector indicating all of the properties (e.g., *is-red*, *can-fly*, etc.) that are both true of the object and relevant to the task. For example, given the input object crow and the task *report category*, the model activates the output unit corresponding to *is-bird* and

no other output units. When given crow with the task *report parts*, the model activates the output units corresponding to *has-wings*, *has-feathers*, and *has-beak*, but not context-irrelevant outputs like *is-bird*. The model is trained in a supervised fashion with the backpropagation algorithm on a large dataset of objects, properties, and tasks derived from human feature norms (De Deyne & Storms, 2008).

In summary, the ISC model demonstrates how learning about the patterns of co-occurrence across objects and their properties using a neural network results in richly structured representations that support a variety of behaviors (see Giallanza et al., 2024 for examples). It further explains how the system accesses sub-components of its knowledge depending on the current task or context: The model learns that certain semantic features are relevant in certain contexts, and it uses a representation of the current task/context (in the context layer) to selectively accesses its knowledge in a way that emphasizes the relevant features (in the context dependent layer).

A major limitation of the ISC model, however, is that it can only do so for contexts it has already experienced and about which it has been instructed—those that are directly labeled as inputs in the environment. For example, the model learns to warp representations to emphasize color information in the *report color* context, but this requires experiencing that context throughout development (learning) and processing an explicit instruction to report color information, implemented by activation of the corresponding task unit. The ISC model lacks the ability to infer on its own that color is relevant in a new context. In the next section, we describe an extension of the ISC model that provides this ability.

A Model of Context Inference

The ISC-CI model (Figure 1b) extends the ISC model by introducing a mechanism for context inference, based on the key assumption that temporal co-occurrence provides a useful basis for inferring shared context. Specifically, it assumes that (1) objects occurring together in

a given context tend to share the properties elicited by that context; (2) these co-occurrence statistics are learned over the course of development; and (3) this implicit knowledge provides a basis for inferring, from a few examples of objects encountered in a new context, both which features are relevant in that context and what other objects are likely to occur in that context. To make these ideas clear, consider the contexts in which you might encounter different kinds of birds: a bird-watching field trip in science class, a visit to the bird section of the zoo, and a picture book about birds. Each situation involves multiple types of birds (e.g., robins, crows, and ravens) and exposure to multiple bird-related properties (e.g., *can-fly*, *eats-worms*, *is-bird*) in various combinations. After these experiences, encountering a new context in which birds are relevant birds (e.g., learning that crows and ravens have hollow bones in the bird section of the Natural History museum) is likely to be interpreted as relating specifically to birds and their properties, implying that other birds like robins may also occur in this new context, and that they will share similar properties (e.g., robins also have hollow bones). Conversely, contexts such as a science lesson on aerodynamics, a visit to a flight exhibit at a science museum, and a film on the history of flight are likely to involve multiple types of flying things (e.g., crows, airplanes, and butterflies) and flight-related properties (e.g., *can-fly*, *has-wings*, *seen-in-the-sky*). This suggests that a new context involving flying objects such as crows and airplanes (e.g., learning that crows and airplanes are associated with Bernoulli's principle) likely relates to all things that can fly, implying that other flying things like butterflies may also occur in this new context and, again, share similar properties (e.g., butterflies are also associated with Bernoulli's principle). Thus the properties shared by items encountered in a situation can provide a clue about what the current context is, what properties are currently important, and what other items are likely or unlikely also to be observed.

The central hypothesis embodied by the ISC-CI model is that learning such environmental structure can support future inferences about which features might be relevant in *novel* contexts, based on the distribution of items that co-occur in those contexts. That is, observing that a new context involves a certain set of objects (e.g., both robins and airplanes) provides

evidence that certain features will be context-relevant (e.g., *can-fly* and *has-wings*), but not others (e.g., *lays-eggs*), based on past experience. Importantly, this process is graded and probabilistic rather than absolute, as any given set of objects can co-occur in different contexts at different frequencies. In particular, features that are broadly true of many objects are less likely to be relevant in a new context than features that are true of the more limited set of objects seen in that context (Griffiths et al., 2010; Xu & Tenenbaum, 2007). This is because there is a low likelihood of observing any particular set of objects in a broad context: there are many animals, but few *Corvidae*, so it is more likely that a context involving both crows and ravens relates to *Corvidae* specifically than it is that this context relates to animals in general.

In the remainder of this section, we describe the ISC-CI model's training environment and architecture in greater detail.

Training Environment

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

in that context (e.g., {sparrow, jaguar}) as the *query set* (following terminology from meta-learning; Thrun & Pratt, 1998).

We generated the episodes using the objects and features in the Leuven Concepts Database (De Deyne & Storms, 2008; Storms, 2001; Ruts et al., 2004). That database contains a matrix of binary judgments provided by human raters indicating, for each object-feature pairing, whether or not the object possess the feature (e.g., does a bear weigh more than 100 lbs? Are kangaroos found in zoos?). After removing duplicate features and features that are only true of 2 items or fewer, the dataset contained 293 objects and 385 features. We constructed a set of episodes by uniformly sampling from the set of features with replacement, so that each episode involved one shared semantic feature that defined the associated context. Given this feature, we generated a support set by uniformly sampling two items sharing the feature, and a query set by uniformly sampling one additional item sharing the feature and one that *not* sharing the feature. For example, one episode consisted of the "zoo" context with the support set {zebra, elephant} and the query set {lion, rat}. In this episode the model had to first process {zebra, elephant} to infer that it was in the "zoo" context (ie, activating the *zoo* semantic feature as the important shared property of zebras and elephants in the current context). It then had to process the query set {lion, rat} to infer that lions occur in the context but rats do not.

The model was trained in a supervised fashion to activate the important shared semantic feature for support items encountered in the episode, and to generate a binary yes/no prediction for each object in the query set indicating whether or not that object belongs in the context. Importantly, each episode involved a *single* semantic feature shared by the support set and relevant to the inferred context that served as the target output for the model. That is, even though the support set may have shared many features, only one of these was relevant to the to-be-inferred context in a given episode. For instance, the support set {crow, robin} could have been sampled from the "bird" context (i.e., *is_a_bird a*s the context-relevant shared

property) or the “animal” context (i.e., *is_an_animal a*s the context-relevant shared property. If occurring in the “bird” context, the semantic feature *is_a_bird* received a target of 1 and the feature *is_an_animal* received a target of zero, and vice-versa if the same two items occurred in the “animal” context. This ambiguity encouraged the model to learn a probability distribution over the possible contexts that could be correct given the support set.

Model Architecture and Implementation

We designed the model to infer the context from the objects in the support set and use this to make predictions about objects in the query set. The model’s architecture (Figure 1B) is similar to the ISC model (Figure 1A), with a context independent layer that encodes cross-context information, a context layer that encodes information about which features are relevant in the given context, and a context-dependent layer which selectively encodes context-relevant information. Unlike the ISC model, however, the ISC-CI model does not receive a context label as *input*; instead, it processes the items in the support set and uses this both to generate an internal representation of context and provide a predicted context-specific shared semantic feature label as *output*.

The ISC-CI model makes inferences about the current context using objects in the support set by sequentially observing each object in the support set, encoding each in the context independent layer, and integrating these representations by taking their average. This can be viewed as a very simple form of recurrence that accumulates (by linearly integrating and normalizing) activity across the items in the support set in the context independent layer. For simplicity, Figure 1B depicts an “unrolled” version of this mechanism, in which the objects in the support set appear to be encoded simultaneously in the context independent layer.

A second extension of the ISC model is that the ISC-CI model makes predictions about which items in the query set occur in the current context. It does so by forming a *context dependent* representation that takes into account the context inferred from the support set

together with the context-independent representation of each query set item. The model then uses this context dependent representation to activate the binary yes/no output units indicating the likelihood that the query set item occurs in the current context.

Model parameters. We implemented the three layers of the ISC-CI model following the architecture of the ISC model, using 64 units for the context independent layer, 128 for the context layer, and 128 for the context-dependent layer. We initialized the weights and biases in the context and context-dependent layers using PyTorch defaults (Kaiming normal, or He, initialization; He et al., 2015). For the context independent layer, we copied the incoming weights and unit biases from the ISC model (Giallanza et al., 2024) so that each one-hot input, corresponding to one of the items in the Leuven dataset, elicited a pre-trained distributed pattern of activation over units. In Giallanza et al. 2024, we showed that these distributed representations explain human semantic similarity judgments among the Leuven items with remarkably good precision. After copying the weights and biases into the context-independent layer, these were frozen (i.e., transfer learning; Pan & Yang, 2009), as we found this increased the performance of the model and the stability of training. The model used the ReLU nonlinearity in the hidden layers and was implemented in PyTorch (Paszke et al., 2019). Parameters for the trained model, along with the training data and code used to run the simulations in this article, can be found at *https://github.com/tylergiallanza/IntegratedSemanticsControlContextInference*.

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

the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001 for a total of 500,000 episodes, by which point the loss converged.

Summary

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

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

In the remainder of this article, we examine the model's ability to carry out inductive inference and similarity judgments in greater detail. In each domain, we first show that the ISC-IC model can account for human data collected in prior work about as well as prior models in

the literature — both ones specifically designed to address these phenomena as well as more general large language models (LLMs) — providing a single, integrated account of classic empirical observations across both inductive inference and similarity judgment. We then consider cases in which the ISC-CI model makes predictions that are very different than prior models, and adjudicate the various models using new behavioral experiments that highlight the use of context-dependent processing.

Part 1: Inductive Inference

Overview

In this section we report the ISC-CI model's ability to account for human behavior in inductive inference tasks, which involve determining if a property, response, or class label common to one set of objects does or does not extend to another object. Several influential models propose that inductive inference relates closely to similarity: The more similar objects are to one another, the more likely they are to share properties, responses, and class labels (Bhatia, 2023; Osherson et al., 1990; Sloman, 1993). Such models often provide a strong fit to human data, insofar as their predictions correlate well with human judgments, and they provide a basis for distinguishing high-confidence inferences from low-confidence ones. Study 1 evaluates how well the ISC-CI model explains human behavior in these such tasks, comparing it to a set of similarity-based models using behavioral data collected in prior studies.

While similarity may be important for inductive inference, this can be complicated by the multi-dimensional relationships between objects (Tversky, 1977) and the possibility that different dimensions may be important in different contexts. For example, ravens, robins, skunks, and foxes vary in both their taxonomic categories and their colors. In some cases, taxonomy is more important for making inductive inferences (e.g., learning that robins and ravens have hollow bones), while in other cases color is more important (e.g., learning that the Spanish word "rojo" applies to robins and foxes). In general, the similarities among items in

similarity-only models, like the feature overlap model, are fixed for all contexts, rendering them unable to account for these cases where the relevance of similarity among different dimensions varies by context.

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

The ISC-CI model's ability to both infer and use context to address this problem may help explain sensitivity of inductive inference to both taxonomic and other kinds of structure, all of which are assumed to reflect similarity along potentially different feature dimensions. When a set of items occur together in a task, these induce a representation of the context that reflects the aspects of semantic structure the items tend to share. This context then shapes the representation of query items in the context-dependent layer so that those also sharing the same semantic structure elicit a positive response, while those that do not elicit a negative response. Thus the model preserves reliance on similarity in two respects: (1) the support items share some dimensions of similarity, which are preserved in the context representation; and (2) the context warps the similarities among query item representations in the context-dependent layer. However, because the context representation itself and its influence on the context-dependent representations both depend on the particular items appearing in the support set, the similarities that govern the model's ultimate decision change with context—providing a possible mechanism for understanding phenomena that seem to challenge similarity-based

approaches. Study 2 evaluates this possibility by comparing ISC-CI and other similarity-based models in their ability to consider different dimensions in different contexts.

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

Prior Models of Inductive Inference

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

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

The Feature Overlap Model

The feature overlap model proposes that people judge the strength of a property induction argument by measuring the degree to which the conclusion shares features with the premises.

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

The feature overlap model requires feature vector representations of objects. Bhatia generated these vectors by using the Feature-BERT model (Bhatia & Richie, 2024) to estimate the probability that each of 25,797 features are true for a given object. To maintain parity with the ISC-CI model, which was trained on data derived from human feature norms in the Leuven Concepts Database (De Deyne & Storms, 2008; Ruts et al., 2004; Storms, 2001), we implemented a version of the feature overlap model using the Leuven features[1]. Critically, it is worth noting that, like other feature overlap models, the values of each vector are always used "as is," without any influence of other arguments or any other elements of the context in which a judgment is made. In this respect, they rely on a fixed metric space for all judgments.

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

The Similarity Coverage Model (SCM)

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

The SCM estimates the strength of an argument by taking a weighted average of the similarity and coverage terms (throughout our studies, we use an even weighting between the two terms, though we found in a pilot study that the value of the parameter has only a small effect on the results). The SCM calculates the similarity term by measuring the similarity (typically defined as mean similarity judgments from human participants) between each premise and the conclusion and taking the maximum of these similarities. Next, the SCM calculates the coverage term by first determining the *covering category*, which is the most specific taxonomic category that contains all of the premises and the conclusion (e.g., the covering category for {crows, ravens} → robins is birds, while the covering category for {crows, alligators} → goldfish is animals). It then measures the similarity between the premises and all objects in the covering category, taking the average of these similarities to form the coverage score. For example, the coverage for the argument {crows, ravens} → robins would be determined by calculating the similarity between crows, ravens, and *sparrows*; then calculating

the similarity between crows, ravens, and *storks*; and so on for all possible birds. The coverage term is the average of these similarity scores.

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

The SCM requires, as input, pairwise similarities between all the objects that may occur in an argument. Gathering pairwise similarities from human raters for all of the objects used in our dataset would be prohibitive, so we instead estimated similarity by calculating the cosine similarity between the Leuven feature vectors representing each object. We found in a pilot study that this produced results similar to those that directly use human similarity judgments (see Supplementary Information).

Large Language Models

The final model class we consider is LLMs (Achiam et al., 2023; Brown et al., 2020). LLMs can judge the strength of a property induction argument directly by processing a prompt. Prior work (Bhatia, 2023; Han et al., 2022; 2024) has found that the performance of LLMs on this task is mixed, with models such as DeBERTa (He et al., 2020) and GPT-3.5 failing to capture certain aspects of human inductive reasoning. More recent work (Han et al., 2024), however, showed that GPT-4 provides a fairly close fit to human behavior when given an appropriate

prompt, suggesting that LLMs are a useful benchmark against which to compare psychological theories of induction. We therefore included GPT-3.5 and GPT-4 as comparisons to our model. We presented property induction arguments to the LLMs following the best-performing prompting strategy outlined in Han et al. (2024).

Study 1: Analyses of Semantic Effects in Property Induction

Study 1a: Argument Strength Ratings

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

More specifically, we aggregated human ratings across several prior studies: one experiment from Rips (1975), consisting of 42 single-premise arguments involving mammals; one experiment from Osherson et al. (1990), consisting of 36 two-premise arguments each of which used the conclusion "horses;" two experiments from Bhatia (2023), the first consisting of 300 single-premise arguments and the second consisting of 300 two-premise arguments, both sampled from six categories (birds, fruits, vegetables, clothing, furniture, and vehicles); and one experiment from Han et al. (2024)[2], consisting of 1,168 one- or two-premise arguments from three categories (mammals, birds, and vehicles). In each study, participants were presented with a series of property induction arguments and asked to indicate the strength of each

[2] We limited our consideration to datasets involving "specific" arguments, which involve basic-level categories such as dogs and cats, rather than "general" arguments, which involve a mix of basic-level categories such as dogs as well as superordinate categories such as mammals, because our model was only trained on basic-level categories. In the Discussion of Part 1 we consider how our model may be extended to account for general arguments in future work.

argument on an interval scale (that we rescaled to the range [0,1]). Each argument involved objects from a common taxonomic category (e.g., in one argument the premises and the conclusion were all mammals, while in another argument the premises and conclusion were all vehicles) with one to three premises and a single conclusion. We removed arguments involving objects that do not occur in the Leuven Concepts Database, resulting in a final aggregate dataset of 1,067 arguments.

We generated strength estimates for these arguments using the feature overlap model, the SCM, GPT-3.5, and GPT-4 following the procedures outlined above. We generated strength estimates using the ISC-CI model by including the premises in the support set and the conclusion in the query set and measuring the response (yes/no) label output provided by the model (see Figure 1B). This resulted in a score between 0 and 1 for each model and for humans, indicating the estimated strength of each argument. We then measured the Pearson correlation between each model's argument strength predictions and human argument strength ratings, grouped by dataset (Figure 2). Error bars indicate the 95% confidence interval of the correlations (at the argument level).

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

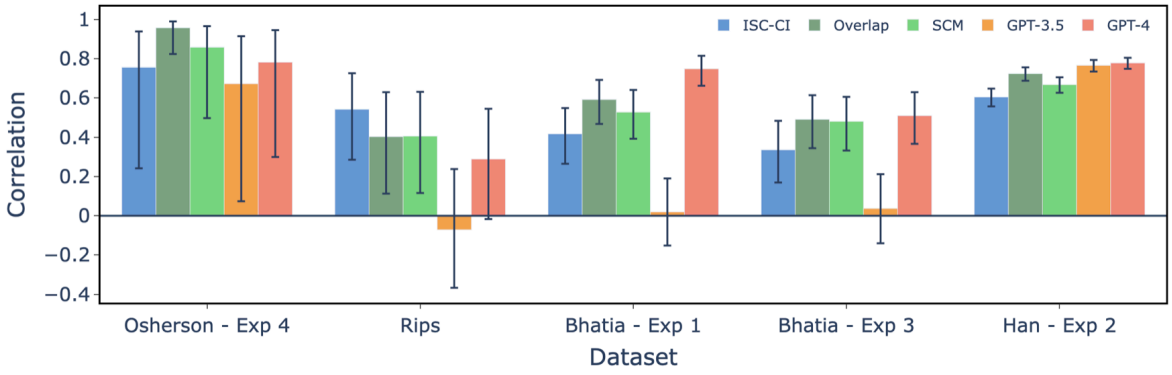

Figure 2. Correlations of model and human performance. Bars indicate the Pearson correlation between each model's argument strength predictions and human argument strength ratings for each of the 5 datasets. Error bars represent the 95% confidence interval of the correlations, measured at the argument level.

Study 1b: Argument Strength Factors

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

Specifically, we focused on five factors from Osherson's (1990) work — some of which align with the similarity-only approach, and some of which challenge it[3]: premise-conclusion similarity, conclusion typicality, premise diversity, monotonicity, and cross-category non-monotonicity. We also evaluated within-category non-monotonicity phenomena initially reported by Medin et al. (2003) and closely related to patterns of name-induction studied by Xu & Tenenbaum (2007) in their Bayesian model of word learning. Thus, in total, we investigated the extent to which phenomena associated with six factors thought to govern human argument-strength ratings, described in more detail below, also arise in simulations with the ISC-CI model, the feature overlap model, the SCM, GPT-3.5, and GPT-4.

---

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

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

Premise-conclusion similarity. As discussed above, people tend to find an argument strong when the conclusion is similar to the premises. For example, the argument {crows} → ravens is stronger than the argument {crows} → robins because ravens are more similar to crows than are robins. Note that this factor depends only on similarity and thus is consistent with similarity-only models. To test this factor, we generated each argument by randomly sampling a category, then randomly sampling a premise and a conclusion from the 10 most typical objects in the category (following Bhatia, 2023). We then calculated the similarity between the premise and the conclusion for each of the arguments (as rated by humans), labeling an argument as strong if its similarity was above the median and weak if it was below the median.

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

rating for objects in that category, the argument was predicted to be strong, otherwise it was predicted to be weak.

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

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

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

either a single premise (for weak arguments) or two premises (for strong arguments) from the 10 most typical objects in the category.

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

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

Cross-category non-monotonicity. Other studies have found non-monotonicity effects when arguments span taxonomic categories: Adding a premise to an argument may weaken that argument when the new premise broadens the lowest-level taxonomic category that includes

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

Results

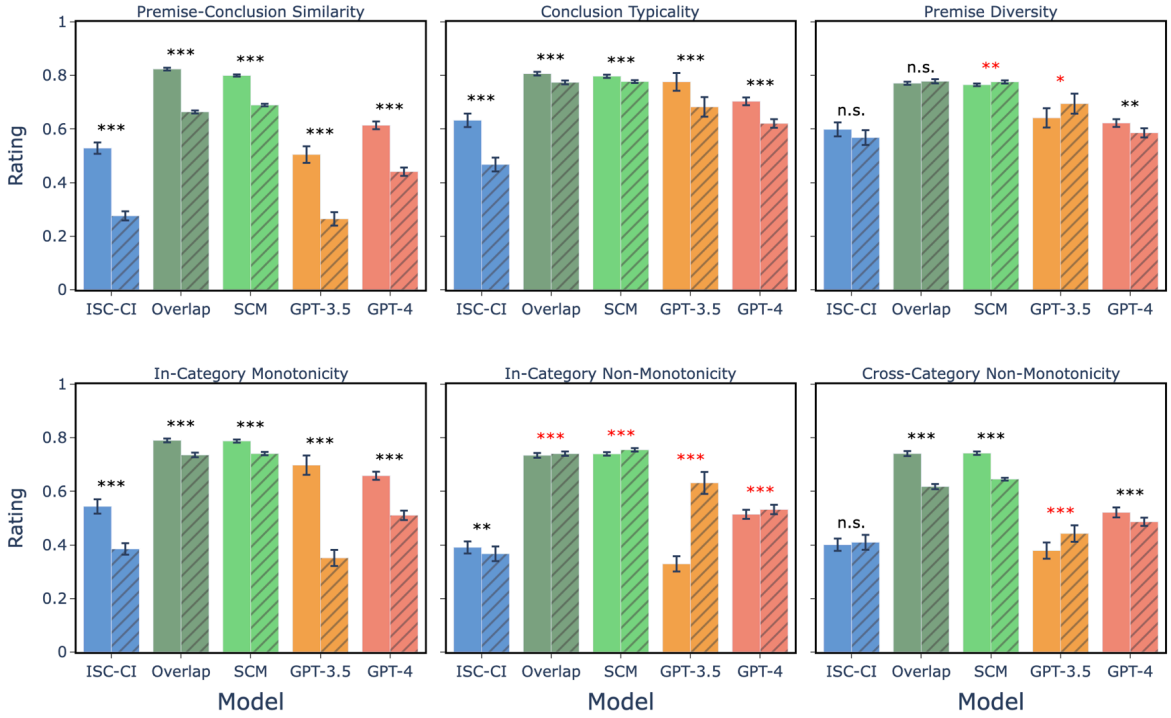

Figure 3. Ratings for strong and weak arguments according to six factors. Bars indicate each model's averaging ratings for the arguments in a given factor, divided into strong (solid shading) and weak (striped shading) groups. Star values above each set of bars indicate p-values from t-tests comparing ratings across the two groups, where n.s. indicates p>=.05, one star indicates p<.05, two stars indicates p<.01, and three stars indicates p<.001. Significant values in the same direction as human ratings are shown in black, while significant values in the opposite direction are shown in red.

Figure 3 shows the performance of the models across the six factors, divided into strength ratings for arguments in the strong and weak groups. Models that show an effect in the predicted direction should provide significantly higher argument ratings for strong arguments (solid bars) than for weak arguments (striped bars). We measured significance with unpaired t-tests for each of the six factors: premise-conclusion similarity, conclusion typicality, and premise diversity, and with paired t-tests for in-category monotonicity, in-category non-monotonicity, and cross-category non-monotonicity. We found that all models showed a significant effect in the predicted direction for premise-conclusion similarity, conclusion typicality, and monotonicity. In contrast, the models diverged in their predictions for premise diversity and the various forms of non-monotonicity, each of which we consider below.

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

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

Non-monotonicity. For in-category non-monotonicity, ISC-CI was the only model to show an effect in the predicted direction; every other model showed an effect in the opposite

direction. In contrast, for cross-category non-monotonicity, ISC-CI showed no effect, while the feature overlap model, the SCM, and GPT-4 all showed an effect in the predicted direction (and GPT-3.5 showed an effect in the opposite direction).

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

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

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

For the ISC-CI model, adding a new premise increases argument strength when it aligns with patterns of coherent covariation among the terms of the argument that give rise to the

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

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

and the conclusion are well-aligned; consider the arguments {sharks} → lions and {sharks, flies} → lions. Here, broadening the context decreases the strength of the argument, because the context induced by sharks alone is more fitting of lions than is the context induced by both sharks and flies. In other words, when there is poor alignment between the first premise and the conclusion, adding a second premise (i.e., broadening the context) likely increases argument strength, but when there is already strong alignment between the first premise and the conclusion, adding a second premise likely decreases argument strength.

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

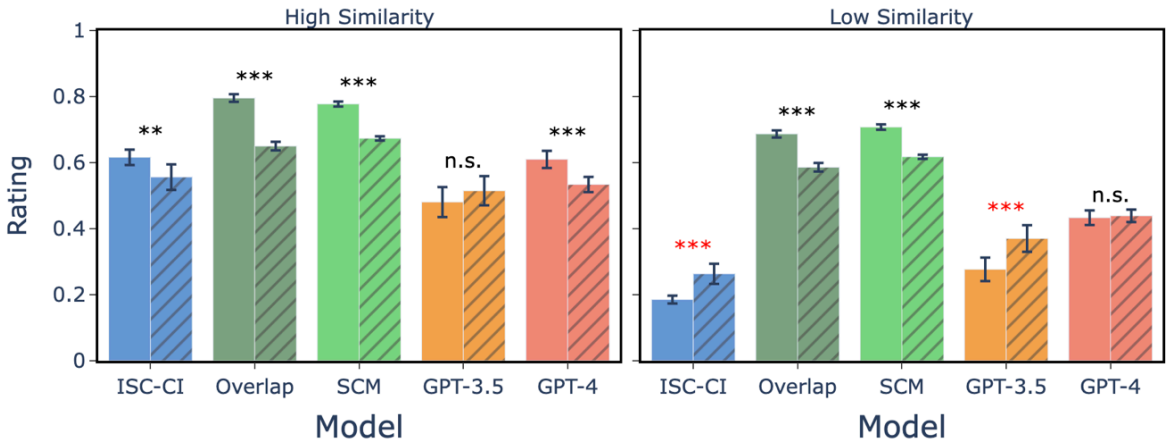

Figure 4. Argument ratings for cross-category non-monotonicity. Bars depict strength ratings for arguments in the cross-category non-monotonicity condition, divided into arguments for which the conclusion is similar to the same-category premise (left) versus arguments for which the conclusion is dissimilar to the same-category premise (right). The primary result is that the ISC-CI model is the only one to demonstrate significant effects in the correct direction across the two conditions.

a significant effect in the high similarity condition, $p<.001$, and no significant effect in the low condition).

In summary, testing in-category and cross-category non-monotonicity effects on a large-scale, ecologically validated dataset reveals interesting differences between the models based on subtle statistical relationships among terms (such as the similarity among the premises and with the conclusion) that may provide a more accurate account of inductive inference than previously described qualitative or discrete factors, such as predefined category relationships and/or taxonomy. We test the extent to which these can explain human judgments in Study 2.

Discussion

Study 1 established that the ISC-CI model provides at least as good an account for a range of prior argument-strength data as prior models, including the SCM designed specifically to account for these data and the feature overlap model shown by Bhatia (2023) to explain many such phenomena. Across 5 studies, predicted strength ratings from ISC-CI correlated significantly with human judgments, with the magnitude of this correlation qualitatively similar to those of the SCM and feature overlap model in all cases. In contrast, predictions of both large language models showed non-significant correlations with human judgments on at least one dataset.

Considering six factors proposed to undergird human ratings of argument strength — concerning semantic relationships among premises and conclusion — the three psychological models performed similarly, showing the predicted effect for four factors and null or reversed effects for two. The ISC-CI model showed null results of premise diversity and cross-category non-monotonicity, with subsequent analyses suggesting that model behaviors depend critically on similarity relations among premises and conclusions in these tasks. The SCM showed reliable effects in the wrong direction for both premise diversity and in-category non-monotonicity—a surprising result given that the coverage mechanism was invoked specifically to explain deviations from pure similarity in such tasks. The feature overlap model

showed no reliable effect of premise diversity, like the ISC-CI model, and a reversed effect for in-category non-monotonicity, like the SCM.

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

Overall, the simulations of property induction demonstrate that all the different modeling formalisms produce comparably good fits to well-known phenomena in the literature, with a few subtle differences. While the subtleties could be worth investigating in future work, one conclusion from these simulations is that the prior empirical record does not cleanly adjudicate the different models. For this reason, Study 2 describes new experiments designed to test explicit predictions of the ISC-CI model that distinguish it from other models.

Study 2: Empirical Tests of Context Effects in Inductive Inference

Rationale

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

A core feature of the ISC-CI model is that the stimuli present in a given setting are combined to generate a context representation that warps the semantic space, shaping how new items are processed by shifting “attention” toward the semantic features or dimensions that they share in common. This can lead to dramatic differences in argument strength ratings

depending on the particular items involved in an inductive inference problem. For example, the items {robin, crow} share the *is-bird* feature in common, while the items {robin, cardinal} share both the *is-bird* and *is-red* features, implying that both features are important for making judgments in that context. The ISC-CI model should therefore show a much stronger rating for the argument {robin, cardinal} → fox than for the argument {robin, crow} → fox, since in the former case the model attends to the color shared by the premises and the conclusion.

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

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

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

Motivated by these considerations, we designed two experiments that manipulated non-taxonomic semantic features in inductive reasoning tasks to test the different predictions made by the SCM, feature overlap, and ISC-CI models. More specifically, drawing on the differences in model performance observed for cross-category non-monotonicity in Study 1, Study 2a tested whether non-monotonicity effects emerge for context-specific categories: Does adding a premise to an argument weaken that argument when the new premise lacks a feature shared by the first premise and the conclusion? For example, the ISC-CI model predicts that the argument {robins} → bees will be rated as stronger than the argument {robins, spiders} → bees, since adding the premise spiders implies that *can-fly* and *has-wings* are no longer as context-relevant as being an insect. In contrast, the feature overlap model and the SCM make the opposite prediction, because spiders are more similar to bees both overall and taxonomically, than are robins. Study 2b directly contrasted taxonomy with context-specific semantic features in a categorization task: Do people actively prefer context-specific categories over taxonomic ones when forced to choose? For example, consider a prompt such as "An unknown category includes sparrows and airplanes. Which is the category more likely

to also include: bats, or penguins?" The ISC-CI model predicts that people will choose bats instead of penguins, because bats, sparrows, and airplanes can all fly, whereas penguins cannot. In contrast, the feature overlap model and the SCM make the opposite prediction because penguins, like sparrows, are all birds but airplanes are not.

Study 2a: Context-Dependent Non-Monotonicity

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

Methods

Stimuli. We generated argument pairs by: 1) first randomly sampling an object from the Leuven dataset that has the feature *can-fly* as the first premise (e.g., robins); 2) identifying an object that also can fly but is from a distinct taxonomic category as the conclusion (e.g., bees); and 3) an additional object from any taxonomic category that *cannot* fly as the second premise (e.g., spiders). This yielded a strong and weak version of each argument (e.g., {robins} → bees vs {robins, spiders} → bees). To better distinguish between the ISC-CI model and the SCM and feature overlap model, we first generated all possible argument pairs using the Leuven stimuli, then selected pairs for which the ISC-CI model chose the single-premise argument while both alternative models chose the two-premise argument, resulting in a set of 80 total argument pairs.

Procedure. At the start of the experiment, each participant was given instructions indicating that they would be asked to make a series of independent decisions concerning the relative strength of arguments. They were then given an example argument demonstrating the structure of the trials that contrasted a single-premise argument with a two-premise argument, such as

"Which of the two arguments is more likely to be true: zebras and donkeys have property X, therefore horses have property X; or zebras have property X, therefore horses have property X?" Each participant then completed a series of 35 decisions, randomly sampled without replacement from the 80 argument pairs, with 5 randomly interleaved attention checks (participants were informed during the instructions that these always had an unambiguous answer, e.g.: {robins} → bees vs {robins} → robins). Participants were asked to make a binary indication of which of the two arguments they felt was stronger on each trial.

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

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

Results

We evaluated the performance of humans and the LLMs by measuring, within participant, the percentage of trials for which the single-premise argument was chosen in favor of the two-premise argument (i.e., the percentage of trials for which non-monotonicity was demonstrated; Figure 5). Consistent with the prediction of the ISC-CI model, we found that the median participant chose the single-premise argument 85% of the time, and 21 out of the 29 participants chose the single-premise argument more often than they chose the two-premise argument. The preference for the single-premise argument was significant by binomial test at both the participant level ($p=.012$) and the argument level ($p<.0001$). In contrast, both LLMs showed a preference for the *two*-premise argument, with GPT-3.5 choosing the single-premise argument only 3.75% of the time and GPT-4 choosing it 1.25% of the time (both significant by argument-level binomial test). Overall, these results support the ISC-CI model's hypothesis that

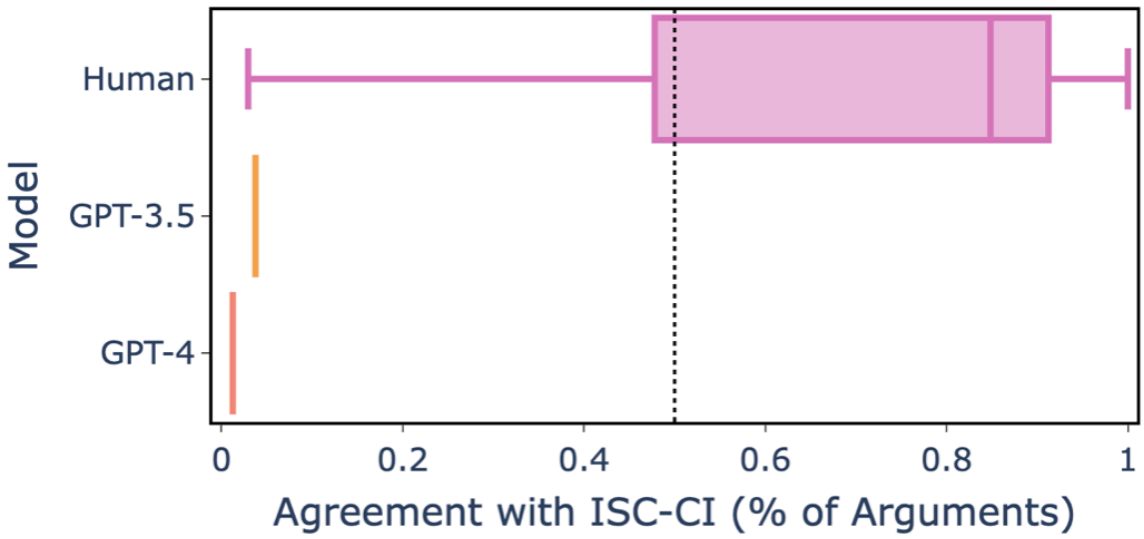

Figure 5. Human and LLM choices for context-dependent non-monotonicity. The x axis indicates the percentage of trials for which humans and the LLMs chose the single-premise argument (agreeing with the ISC-CI model) over the two-premise argument (agreeing with the overlap and SC models). The human data indicate participant-level results, showing the distribution over how often each participant chose the single-premise argument.

non-monotonicity effects extend to non-taxonomic semantic features, effects that are neither predicted by the SCM or feature overlap model, nor exhibited by either of the LLMs.

Study 2b: Context-Dependent Categorization

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

Methods

Stimuli. Using the six context-specific features, we constructed categorization questions by selecting two target objects, one context-specific choice object, and one taxonomic choice

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

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

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

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

Results

We evaluated the performance of humans and the LLMs by measuring, within participant, the percentage of trials for which the context-specific match was chosen in favor of the taxonomic match (Figure 6). If humans have a preference for the context-specific match, as predicted by the ISC-CI model, each participant should choose the context-specific match more often than the taxonomic match. Consistent with this prediction, we found that the median participant chose the context-specific match 84% of the time, and 43 out of 50 participants chose the context-specific match more often than they chose the taxonomic match. The preference for the context-specific match was significant by binomial test at both the participant level, $p<.0001$, and the argument level, $p<.0001$. In alignment with human ratings, GPT-4 chose the context-specific match 71% of the time (significant by argument-level binomial test, $p<.0001$), while, in contrast, GPT-3.5 chose the context-specific match only 24% of the time (significantly in favor of the *taxonomic* match by argument-level binomial test; $p<.0001$). Overall, these results demonstrate that people can categorize objects on the basis of less prominent semantic features, rather than the generally more prominent taxonomic categories, when those features are shared by items encountered in a given context (and that the same behavior arises in state-of-the-art LLMs).

Discussion

Study 2 demonstrates how the central mechanism in the ISC-CI framework—inferring a context representation from blended representations of the items encountered in the context—can explain and predict patterns of human behavior in inductive inference tasks. After learning,

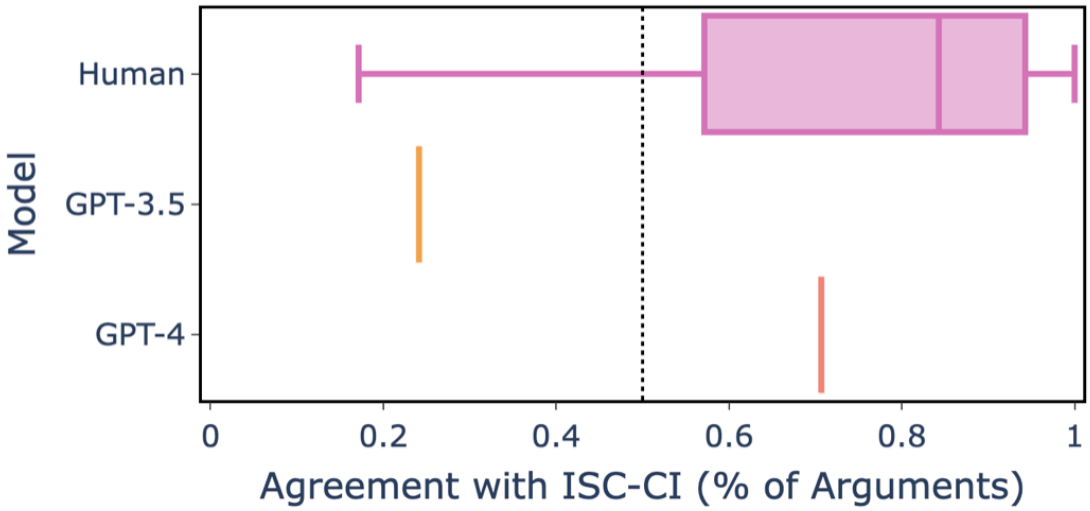

Figure 6. Human and LLM choices for taxonomic vs context-specific categorization. The x axis indicates the percentage of trials for which humans and the LLMs chose the context-specific match (agreeing with the ISC-CI model) over the taxonomic match (agreeing with the overlap and SC models). The human data indicate participant-level results, showing the distribution over how often each participant chose the context-specific match.

the model can generate context representations elicited by a novel combination of objects (premises) to identify the properties they share, and use this to shape processing of another object (a conclusion) by “directing attention" to dimensions along which the premises are most similar. We showed that this mechanism produces responses that closely correspond to those of humans in both property induction and categorization judgments.

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

infer that color-based relationships matter more in the second argument than in the first, and that robins and foxes are more similar in the context of the second argument than they are in the context of the first argument.

The feature overlap model suffers from a similar limitation: It treats all features as equally important regardless of the particular premises. This means that its argument strength scores are always dominated by the *number* of features shared by the premises and the conclusion, not by the *types* of features that are shared[4]. The arguments {robin, crow} → fox and {robin, cardinal} → fox thus receive very similar scores because only a single additional feature (*is-red*) overlaps in the second argument compared to the first argument. Since objects belonging to the same taxonomic category tend to share large numbers of features in common, taxonomy exerts a strong influence on the feature overlap model's processing across all arguments.

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

---

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

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

Summary of Part 1. In the introduction we emphasized a critical tension in theories of human inductive inference: while semantic similarity appears to explain many phenomena, a range of classic results has suggested that similarity alone is not sufficient to adequately explain human behavior. Alternative models thus invoke additional representational structure to explain such phenomena, such as discrete categories situated within a taxonomic hierarchy, which are used to compute critical information during inference such as the "coverage" term in the SCM (Osherson et al., 1990), or the prior probabilities in Bayesian approaches (Xu & Tenenbaum, 2007). We proposed an alternative view, in which systems of semantic representation and control are integrated within the same representational space shaped by statistical learning, so that representations of contextual information can *reshape* the semantic similarity structures deployed in a given task or trial. This view reconciles "similarity-only" and "similarity-plus" theories by proposing that: (a) inference judgments are always shaped by

semantic similarities within some representational space; but (b) the space can be warped by representations of context to amplify task-relevant semantic structure and minimize task-irrelevant structure.

The studies reported in this section compare a model implementation of this proposal (ISC-CI) to other models in the literature, with three important conclusions. First, without any task-specific tuning or emphasis, the ISC-CI model accounts about as well as other models for data from prior studies of human inductive inference. Second, as Bhatia (2023) has also argued, prior work may have mischaracterized the limitations of similarity-based models. In our simulations, a pure similarity (feature overlap) model provided as good a fit to human judgments as did the similarity-coverage model, even for phenomena thought to demonstrate the need for additional mechanisms beyond similarity (e.g. conclusion typicality, cross-category non-monotonicity). Moreover, a key phenomenon motivating the similarity-coverage proposal—premise diversity—was not actually captured by the SCM, which in fact showed the reverse pattern. In general, Study 1 suggested that prior results do not strongly differentiate the various models when these are fit to a large and representative corpus of semantic feature norms.

Third and critically, the ISC-CI model suggests one important way that semantic similarity, either alone as in the feature overlap model or together with "coverage" information in the SCM, is insufficient to explain human inductive inference. Specifically, the model predicts that people can discern elements of semantic relatedness amongst a collection of premise items even when these cross-cut taxonomic or other generally important elements of similarity, and can direct "attention" to this common structure in order to judge whether novel items do or do not share the same context-relevant properties. Study 2 provided strong evidence that people behave as predicted by the ISC-CI model, and in contradiction to the predictions of both overlap and SC models, in such cases.

The patterns of inference we have considered, however, provide only part of the historical case challenging similarity as a construct for understanding induction. A remaining part

concerns studies of similarity judgments themselves, which sometimes reveal puzzling properties that challenge the coherency of similarity as an explanatory principle. We turn to these phenomena in Part 2.

Part 2: Similarity

Overview

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

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

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

In this section we consider how the ISC-CI model can address these phenomena, through the same use of context-modulated distance that allowed it to address the phenomena associated with inductive inference considered above. This provides a reconciliation of traditional distance-based views with the context-dependencies identified by Tversky's work That is, like standard distance models, our approach assumes that objects have vector representations and that distances between these representations determines similarity judgments. However, in contrast to standard distance models, the ISC-CI model constructs and uses representations of context to warp these distances by differentially weighting different dimensions based on the current context, thereby shaping the similarity relations that drive similarity judgments. Accordingly, we show that the same mechanisms accounting for patterns of induction in Part 1 allow the ISC-CI model to account for both standard pairwise similarity judgments (Study 3), which are well-explained by distance-based models, and the specific biases observed by Tversky (Study 4).

Models of Similarity

We compared the ISC-CI model to two prior models of similarity judgments from the psychology literature: the feature overlap model (representative of standard distance-based models; Sloman, 1993) and the feature contrast model (representative of Tversky's approach; Tversky, 1977). As discussed in Part 1, the feature overlap model proposes that objects are similar to the extent that they share features with one another, which can be calculated by representing each object with a feature vector and measuring the cosine similarity as a measure of the distance between the two vectors.

The feature contrast model also builds on the idea that objects are more similar when they share features in common and less similar when they possess unique features. However, it does so with set representations rather than vectorial ones. This allows for greater flexibility relative to feature overlap model, by dissociating shared features from features unique to one of the objects. Specifically, the model represents each object *a* as a the set of binary features *A* that are true of that object. The intersection of two set representations, $A \cap B$, contains the features that the objects share, while the differences between the sets, $A - B$ and $B - A$, contain the features unique to one of the objects. A weighted combination of the cardinality of these sets provides a measure of object similarity: The similarity between the objects *a* and *b* is $S(a, b) = \theta\,|A \cap B| - \alpha\,|A - B| - \beta\,|B - A|$, with scalar weight parameters $\theta, \alpha, \beta \geq 0$. Tversky showed that modulating these weight parameters in different contexts can account for some of the biases in human judgments that cannot be captured by distance based models. For example, Tversky (1977; Tversky & Gati, 1978) showed that setting $\alpha \neq \beta$ produces asymmetric similarity judgments in the same direction as humans. We set the parameters following Tversky for the simulations in Study 3 and Study 4.

The ISC-CI model proposes that objects are similar to the extent that they co-occur with one another across different contexts. We can measure this by combining two factors: how often item B occurs in contexts that involve item A, and how often item A occurs in contexts that involve item B. Analogous to the methods in Part 1, we can compute how often B occurs in contexts involving A by using A as the *premise item* and B as the *query item*. The strength of the "yes" response unit then provides a numerical measure of co-occurrence. Like the feature contrast model, combining these two terms allows ISC-CI the flexibility to produce different similarity judgments depending on how the question is framed, for example by producing different similarity scores for the questions "how similar is A to B?" and "how similar is B to A?". When similarity judgments are presented bidirectionally, such as "how similar are A and B

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

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

Study 3: In-Category Similarity

Rationale

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

Methods

We tested how well the models account for the human similarity judgments provided in the Leuven Concepts Database. These consist of pairwise, bidirectional similarity ratings between all pairs of objects within the same superordinate category. Each pair was presented to between 2 and 4 participants, and each participant judged between 15 and 25 objects, indicating the similarity between the objects on a scale of 1 to 20. We normalized the similarity ratings within-participant by subtracting from each rating the participant's mean rating and dividing by the standard deviation of the participant's ratings, resulting in z-scored ratings for each participant. We then measured cross-subject reliability by correlating each participant's normalized ratings with the average normalized ratings. Finally, we simulated similarity judgments using each of the models, calculated the Pearson correlation between the models'

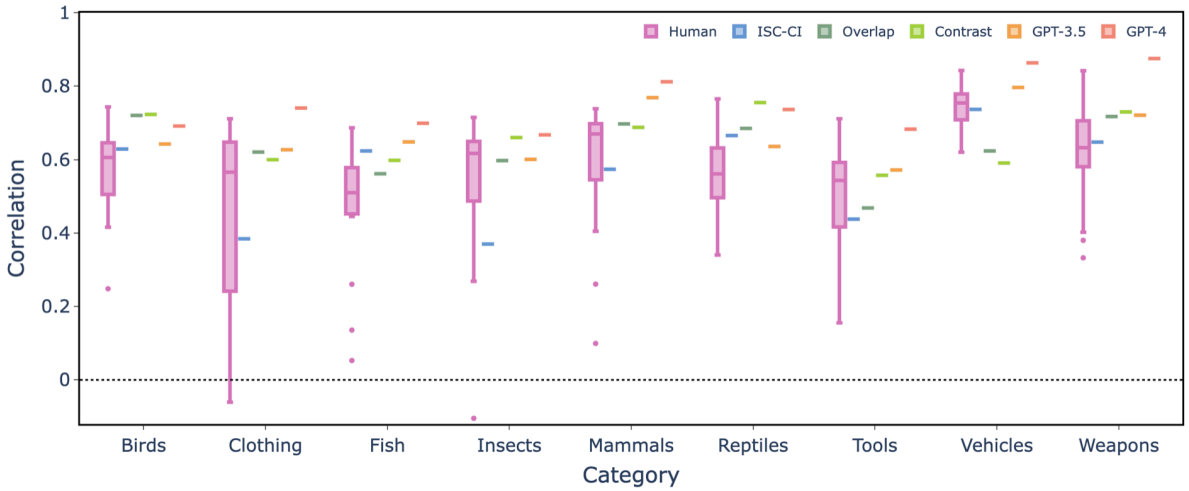

Figure 7. Pearson correlations between model and human similarity scores. Correlations are between each model's predicted similarity scores and the average similarity score reported by participants for each pair of objects within each domain. The box plot for human responses (pink) represent cross-subject reliability, showing a distribution over how strongly each individual participant's responses correlate with the average response across participants.

similarity judgments and the normalized human similarity ratings averaged across participants (for the feature contrast model, we used the parameters $\theta = 1, \alpha = 0.5, \beta = 0.5$, though the results were robust across a reasonable range of parameter values), and used cross-participant reliability to determine the significance of the relationship between each model and human performance.

Results

Figure 7 shows the results, grouped by category. Across all categories, the ISC-CI model's similarity ratings correlate fairly well with human similarity: they are above the lower fence of the human reliability for all categories, and within or above the inter-quartile range for every category except insects. The other models also correlated well with human similarity ratings, showing a qualitatively similar pattern to the ISC-CI model. Overall, the results indicate that both the ISC-CI model and all other models provide a reasonable fit to human standard similarity judgments, and thus that such judgments are not useful for adjudicating among them.

Study 4: Context Effects in Similarity Judgments

Rationale

In this study, we focused on two cases in which human similarity judgments have been reported to contradict axioms of metric spaces (Tversky, 1977; Tversky & Gati, 1978). In one, Tversky & Gati (1978) showed that people can exhibit asymmetry in directed similarity judgments; for example, producing different ratings for "North Korea is like China" than for "China is like North Korea." If similarity judgments are governed by a fixed distance between the two items within a metric space, participants should give the same answer regardless of the direction of comparison. Second, they showed that in multi-alternative similarity tasks people can make decisions that imply very different distances between the same two items. For instance, when asked to decide whether England, Iran, or Syria is the most similar to Israel, people typically choose England, suggesting it is closer to Israel than is Iran. However, when asked to decide whether England, Iran, or *France* is the most similar to Israel, people choose Iran—suggesting it is closer to Israel than is England. This discrepancy could not arise if, in making their decisions, people are consulting fixed distances within a common metric representational space.

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

include the feature overlap model since, by definition, it cannot generate asymmetric or mutually-inconsistent judgments across the conditions of interest.

Study 4a: Asymmetry

Rationale

We first considered asymmetry in directed similarity judgments: the finding that people generate consistently higher ratings between items when they are presented in one order versus the reverse order (e.g., people generally rate “donkeys are like horses” with a high similarity score but “horses are like donkeys” with a lower similarity score). In the ISC-CI model, the first item in the comparison can be viewed as the *premise* item, and the second can be viewed as the *query* item. The premise item provides a context in which the similarity of the query item is judged, and the two different orderings can be simulated by manipulating which item serves as the premise and which as the query. Similarity ratings were taken to be proportional to the activity over the response units: the more similar the query item was to the premise, the more activation should accrue on the *yes* response unit.

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

Methods

We began by finding pairs of objects in the Leuven Concepts Database that yield reliably asymmetric similarity judgments, complementing (and extending) prior empirical data for directed judgments that have generally been restricted to pairs of countries (Aguilar & Medin, 1999; Johannesson, 2000; Tversky, 1977; Tversky & Gati, 1978), narrative stories (Bowdle & Gentner, 1997), or non-literal similes (Ortony et al., 1985). We first chose the 50 object pairs in the Leuven Concepts Database that exhibited the greatest bidirectional similarity with one another (as measured by bidirectional human similarity judgments), then used these in a behavioral experiment in which participants were presented with two directional similarity statements involving the same pair of objects and asked to make a binary decision about which statement seems stronger. For example, on a given trial a participant might be asked: "Which of the two statements seems stronger — synthesizers are like pianos, or pianos are like synthesizers?" Experimental instructions emphasized that there was no correct answer to the questions and that the statement should be chosen that "seems stronger." Each participant completed 35 such judgments, with 5 randomly interleaved attention checks that had an unambiguous answer (e.g., which of the two statements seems stronger: horses are like zebras, or horses are like horses?). We collected data from 55 participants on Prolific, excluding five due to failed attention checks. The study was approved by the Princeton Internal Review Board (Protocol 6079).

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

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

Results

We found that the ISC-CI model, the feature contrast model, and GPT-4 all reliably produced asymmetric judgments in the same direction as humans (Figure 8), agreeing with humans on 76%, 88%, and 85% of the 34 pairs, respectively (all significant by pair-level binomial test; ISC-CI: $p=.0014$; feature contrast: $p<.0001$; GPT-4: $p<.0001$). In contrast, GPT-3.5 performed at chance levels, agreeing with humans 56% of the time. We next tested how well the models account for the magnitude of the asymmetry effect by regressing the models' asymmetry scores against the human asymmetry scores. We found that both the ISC-CI and feature contrast models predicted the magnitude of asymmetry at well above chance levels, with R^2 scores of 0.31 and 0.60, respectively (Figure 8; GPT-3.5 and GPT-4 were excluded from this analysis given that they produced binary, rather than scalar, predictions for each argument pair). Taken together, the directional and regression results demonstrate that both the ISC-CI and feature contrast models can account for asymmetry in similarity judgments.

Study 4b: Multi-Alternative Context Effects

Rationale

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

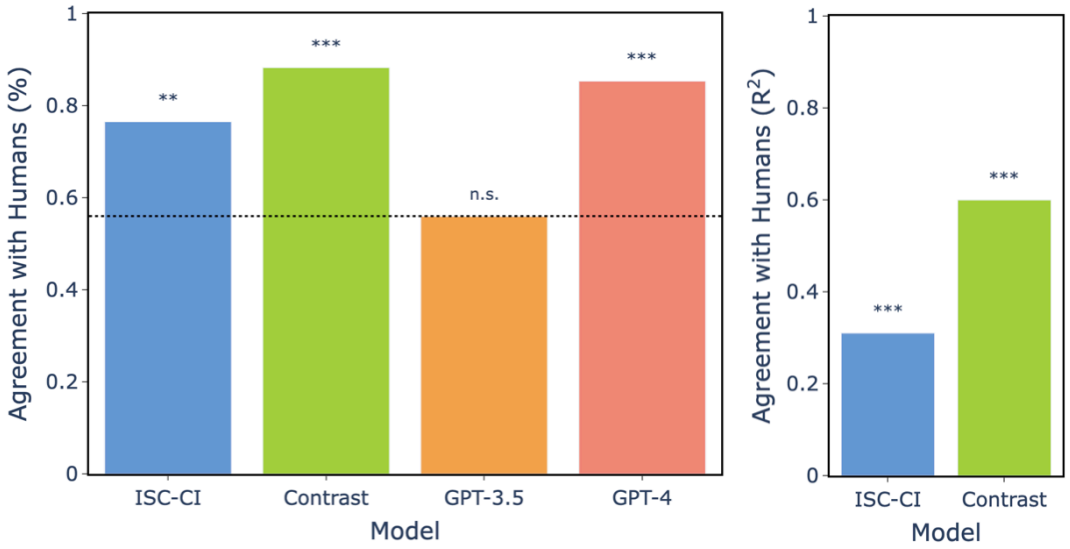

Figure 8. Agreement between model and human asymmetries in directed similarity judgments. *Left:* The percentage of pairs for which each model's directional preference matches human directional preferences. *Right:* Correlations between the magnitude of the asymmetry scores produced by the ISC-CI and feature contrast models and human asymmetry scores.

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

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

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

Methods

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

For incorporation into the ISC-CI, we implemented a variant of the Usher & McClelland (2004) model that replaces the random attentional selection mechanism with the model's

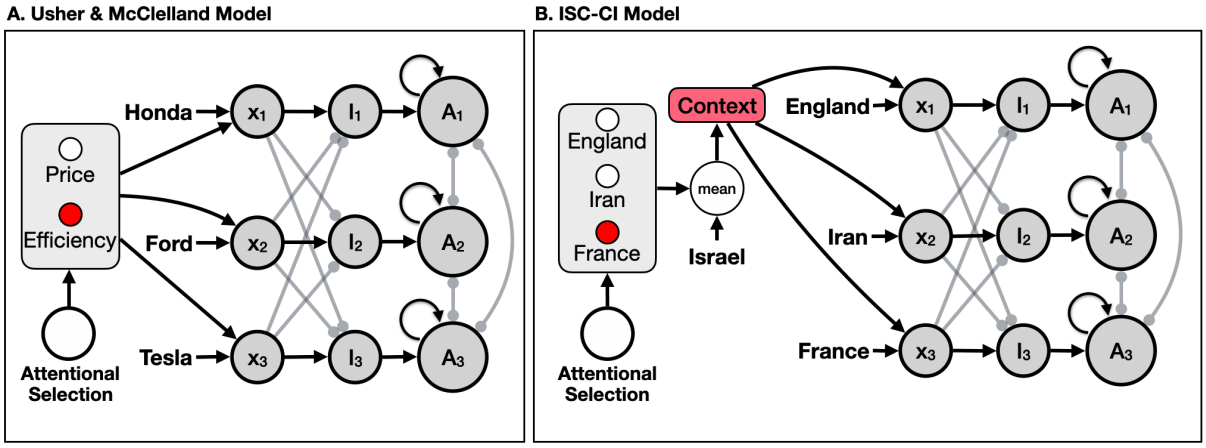

Figure 9. Models of multi-alternative choice. A. Usher & McClelland's (2004) model of multi-alternative multi-attribute value-based decision making In the example, the decision maker is choosing between purchasing a Honda, Ford, and Tesla automobile, each of which varies in price and fuel efficiency. At each time step of processing, an attentional selection mechanism randomly decides between price and efficiency, and the relative value of each option along that dimension is accumulated, with the decision determined by which of the options (A_i) is most active after a fix amount of time has elapsed (or, alternatively, one of the options either crosses a fixed activity threshold or prevails over the others by some margin). B. The ISC-CI model of multi-alternative similarity judgments. Here, the decision maker is choosing which of England, Iran, or France is most similar to Israel. On each time step the attentional selection mechanism randomly selects between the context formed by the target (Israel) and one of the three countries, and that context is used to determine the similarity between the target and each of the items, which is then passed to the decision making mechanism.

context inference mechanism to determine feature attentional weightings in the input (Figure 9). Rather than randomly selecting a semantic *feature* on each time step, the model randomly selects a *context* formed by the target item and one of the choice items. For example, when deciding if England, Iran, or France is the most similar to Israel, the ISC-CI model switches between three contexts: first, the context formed by Israel and England; second, the context formed by Israel and Iran; and finally, the context formed by Israel and France. In each of these three contexts the model measures the similarity between the target and each of the options with respect to that context. For example, in the Israel/England context, the model estimates the similarity between Israel and England, Israel and Iran, and Israel and France. The model then takes the softmax over the context-dependent similarity scores and presents these as inputs to the input preprocessing stage, and then to the LCA decision stage.

Finally, we made one additional change to the Usher & McClelland (2004) model, by using a different nonlinearity when evaluating the differences between the option values. Their model used a non-linearity that included loss aversion because they were modeling value-based decision making, in which people demonstrate loss aversion (Khaneman & Tversky, 1979). Since our model simulates similarity and not value-based judgments, instead we used a simple rectified linear function (i.e., ReLU) that did not include any value-based bias.

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

To test the model empirically, we designed an experiment to measure human-like context effects in multi-alternative similarity judgments using objects from the Leuven data set. Following Tverksy & Gati (1978), we generated paired sets of objects to test how changing one of the options affects choice preferences between the other two options. We denoted a pair of

sets as {*t, a, b,* c_a} and {*t, a, b,* c_b}, where the sets differed only in their final object. Each set was used to evaluate which of the three choice options (*a, b* or *c*) was judged to be most similar to the target object *t*. We designed the sets such that people were likely to choose object *a* when the choice set contained c_a and object *b* when the choice set contained c_b. For example, one set pair consisted of blackbirds as the target and either the options {mice, airplanes, buses} or {mice, airplanes, hamsters}. Consistent with Tversky's initial intuition, the former set creates distinct clusters by *animacy* as the relevant feature, placing "mice" closer to the target, whereas the latter set creates distinct clusters by *flight* as the relevant feature, placing airplane closer to the target. Thus, we anticipated that people should judge mice more similar to blackbirds than airplanes in the first set, but should make the reverse decision in the second set. We predicted that our model of similarity judgment should show the same tendency.

We generated the set pairs using a multi-step process. First, we sampled a large number of object sets {*t, a, b,* d_a*,* d_b} such that all of the following object pairs were similar to one another along distinct semantic dimensions: *t* and *a*, *t* and *b*, *a* and d_b, and *b* and d_a. For example, blackbirds (*t*) and mice (*a*) are similar because they are both animals; blackbirds (*t*) and airplanes (*b*) are similar because both can fly; mice (*a*) and hamsters (c_b) are similar because they are both rodents; and airplanes (*b*) and buses (c_a) are similar because they are both vehicles. Second, we selected from the candidate object sets those object sets that produced context effects using Tversky & Gati's (1978) clustering method (see Supplementary Information for a more detailed procedure). This resulted in 41 object sets that we used to construct the experimental stimuli.

We presented human participants with the generated object sets and elicited similarity judgments. On each trial of the experiment, participants were presented with a multi-alternative similarity question of the form "Which of the following is the most similar to blackbirds: mice, airplanes, or buses?" Each participant completed 35 similarity judgments, randomly sampled

from the 41 object sets. Each participant either saw the version of the object set that included c_a or the version that included c_b. We collected data from 53 participants on Prolific, excluding 3 due to failed attention checks. The study was approved by the Princeton Internal Review Board (Protocol 6079).

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

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

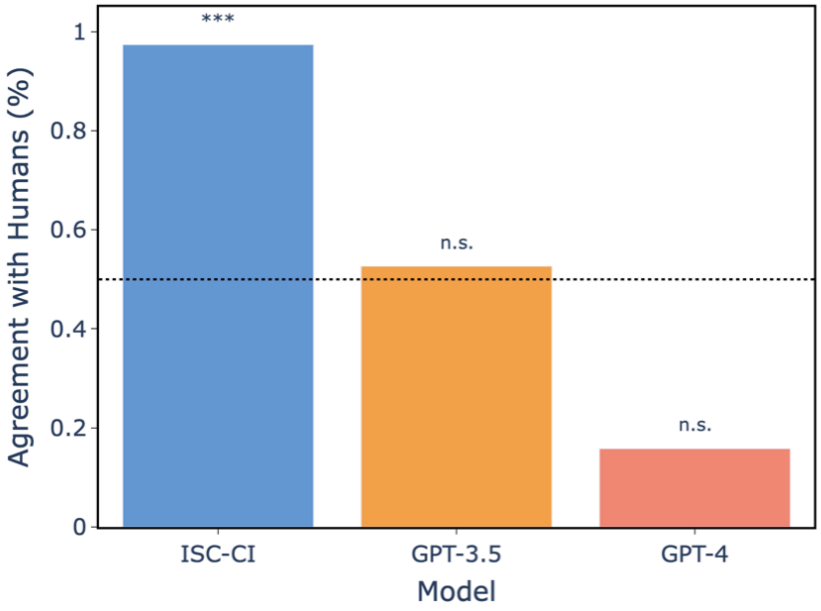

Figure 10. Comparison of model and human choices in multi-alternative similarity judgments. Agreement represents the percentage of object sets for which each model produces context effects in the same direction as humans.

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

Results

As shown in Figure 10, the model predictions agreed with human decisions for 36 out of the 38 object sets, reliably better than chance responding by binomial test ($p<.0001$). GPT-3.5 agreed with human decisions on 21 of the 38 object sets, not reliably better than chance, whereas GPT-4 agreed with human decisions on only 7 of the object sets—reliably *worse* than chance by binomial test ($p < 0.0001$).

Discussion

Studies 3 and 4 demonstrate how the ISC-CI model explains human similarity judgments both when these are consistent with standard distance-based models and when they disagree with such models. For tasks that minimize the effect of context, such as pairwise bidirectional similarity rating (how similar are horse and zebra to one another?), judgments rely on context-independent similarities that reflect the overall covariance structure of the environment, as in standard similarity-based models. For tasks in which context may influence the judgments to be made—for instance, directional comparisons (how similar is a zebra to a horse?), or multi-option arrays (which of these options is most similar to a horse?)—the model's context-inference mechanism can establish which semantic dimensions are relevant to the immediate task, reshaping the similarities that ultimately drive the judgment. Thus the same mechanism that explains patterns of inductive inference in Part 1 also resolves the seeming discrepancy between distanced-based and other approaches to modeling human similarity judgments.

The ISC-CI can be seen as integrating critical features of each of the earlier models. With respect to the feature overlap model, it shares the fundamental assumption that semantic

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

The feature contrast model complements the feature overlap model, by incorporating the flexibility of context-sensitivity, but applying this over a non-parametric (set-theoretic) computation of similarity. This is sufficient to account for asymmetries in directional comparisons under the assumption that people pay more attention to the features of object *a* than to the features of object *b* in a directed statement of the form "*a* is like *b*". Given this differential attention, the model produces asymmetric judgments whenever $|A - B| \neq |B - A|$, where $|A - B|$ is the magnitude of the set of features unique to object *a*. The contrast model only makes pairwise similarity judgments, however, so it cannot easily account for multi-alternative choices without additional assumptions. Tversky & Gati (1978) provided one such assumption: People selectively attend to "diagnostic" features that help cluster objects into smaller groups. Although the diagnosticity principle is intuitively appealing, it does not explain how people determine which features have diagnostic value in any given setting (but see Kruschke, 1992; Nosofsky, 1986; 2011 for potential approaches), which becomes particularly complicated for high-dimensional inputs, nor does it explain why clustering is necessary in multi-alternative choice. The feature contrast model therefore cannot provide quantitative predictions about multi-alternative similarity judgments.

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

Finally, both GPT-3.5 and GPT-4 provided strong fits to bidirectional pairwise similarity judgments. GPT-4 was also able to account for asymmetries in directed similarity judgments, but it was unable to account for context effects in multi-alternative choices, which was surprising given its strong performance in the other studies.

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

General Discussion

In this article, we have presented a model of how people infer which parts of their semantic knowledge are relevant in a given context, how they use that knowledge to guide their

inferences and decisions, and how these capabilities emerge through statistical learning; that is, learning about the co-occurrences between objects, their semantic features, and the contexts in which these are used. The model extends the integrated semantics and control (ISC) model (Giallanza et al., 2024) — which suggests how semantic representations formed from such statistical learning can both shape and be shaped by systems that support attention and control — to provide a means for inferring which semantic features and which objects might be relevant in a given context, and directing attention to those features when making inductive inferences and judging object similarity. Critically, we showed in simulations and experiments that these ideas can resolve a long-standing tension between similarity-based approaches to induction (Bhatia, 2023; Sloman, 1993) and a variety of classic findings suggesting that context plays a critical role in both similarity judgements and inductive inference (Tversky, 1977; Tversky & Gati, 1978). On the one hand, the ISC-CI can be viewed as similar to other models that assume that both of these rely on the relative distances among representations in a metric space. However, it differs from previous models, which assume a *fixed* representational space, by allowing the space to be warped by context representations. This was a central tenet of the ISC model. Here, we show that this provides a mechanism by which inductive inference and similarity judgments can be impacted by context, similar in spirit to classic models that have been proposed to account for effects that can't be explained by similarity-based models that assume a fixed representational space. At the same time, unlike previous models that invoke the influence of context in terms of intuitive accessible but pre-specified, discrete, qualitative factors, the ISC framework grounds this in terms of quantitatively specified context representations, that arise from the same statistical learning mechanisms as the underlying semantic representations over which they preside. Based on this approach, the ISC-CI model accounted not only for standard findings about the factors that have been take the reflect the influence of context on inductive inference and similarity, but also for novel, subtle context effects in both domains that are difficult to explain using prior models.

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

Co-Occurrence, Coherent Covariation, and Control

Temporal Structure and Co-Occurrence

One of the key differences between the ISC-CI model and prior work on semantic cognition is the temporal structure of the environment used to train the model. Neural network models are typically trained in temporally *interleaved* environments, in which objects and their properties are presented in a random order. Interleaved training is motivated by the widely recognized problem of catastrophic interference (McCloskey & Cohen, 1989): Repeatedly presenting a partially trained network with a new fact (e.g., penguins are birds that cannot fly) can interfere with existing knowledge (e.g., most birds can fly), wherein the network begins to extend idiosyncratic properties of the new fact (*cannot-fly)* to all of the other objects in the dataset (i.e., forgetting that most birds can fly). Interleaving presentation of new information with other examples from the dataset (e.g., robins and sparrows, that can fly) reduces or eliminates this interference. Interleaved training has therefore become standard practice for neural network models of semantic cognition (following McClelland et al., 1995).

In contrast, the ISC-CI model is trained with a temporally *blocked* environment, in which objects and properties are presented in a series of contexts each of which involves a set of

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

The ISC-CI model's training environment comports with the temporal structure of events that has been observed in developmental populations (Slone et al., 2023), where objects sharing features with one another tend to cluster together in time. However, the training environment greatly simplifies the complexity a human learner is faced with by segmenting continuous experience into a set of discrete training blocks, each of which corresponds to a distinct context. Incorporating such segmentation directly into the model, rather than pre-specifying it in the training data, is an important challenge for future work. One possibility, in keeping with the broader philosophy of the ISC framework, is that discrete segmentation into distinct contexts is a simplifying assumption that imposes too restrictive a structure on temporal co-occurrence. Instead, recent models suggest that temporal structure itself is subject to continuous, graded similarity between contexts that can be explained again by the same principle of coherent covariation underlying learning in the ISC and ISC-CI models (Giallanza et al., 2024b; Schapiro et al., 2013). Bridging between the context inference mechanism in the ISC-CI model and the temporal structure learning mechanisms in these recent models is an important challenge for future work.

In addition to mitigation of catastrophic interference, an important consequence of the blocked environment is that it exposes the model to temporal co-occurrence. Each context involves multiple objects that occur at the same time; the model learns to invert this process by

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

The ISC-CI model extends these effects of co-occurrence and coherent covariation to higher levels of representation (i.e., of *contexts*) by using the integration mechanism in the context layer to learn about the co-occurrences between objects and other *objects* within and across different contexts. This results in representations that express specific contexts, as well as relationships between the contexts, representing similar contexts — meaning contexts that involve the co-occurrence of similar objects — using similar patterns of activity in the context layer of the network. These representations in turn support pattern completion of a different kind: After learning that a new context involves both robins and ravens, the context will be represented with a pattern of activity similar to other contexts that also involve birds, which supports the inferences that: a) the new context involves bird-like features such as *is-bird* and *can-fly;* and b) the new context involves other birds such as bluejays and ravens.

Structured Context Representations and Cognitive Control

The ISC-CI model extends work using the ISC model that addresses the relationship between semantics and cognitive control (Giallanza et al., 2024). Cognitive control involves the use of *control representations*, which encode information about the current task, goal, or

context, to guide processing of the current stimulus by selectively emphasizing context-relevant semantic features for both inference and response selection. As in the ISC model, the representations in the context layer of the ISC-CI model act as control representations by shaping processing in the context-dependent layer of the network. The versatility and complexity of these representations extends that of both classic models of cognitive control (e.g., Cohen et al., 1990; Miller & Cohen, 2001) and the ISC model (e.g., Giallanza et al., 2024) model in two important ways.

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

Context Inference. Second, most prior models of cognitive control require an explicit signal or cue that indicates the current task. This pre-supposes that the system already knows all the relevant information about the current task, context, or goal ahead of time. However, this is generally not the case when making inductive inferences or similarity judgments, which may involve novel combinations of stimuli, and therefore contexts. The ISC model showed that

online adaptive mechanisms can help shape context representations to optimize processing for a given task (e.g., like human participants, after observing that all of the experimental stimuli involve animals, it can shift its context representation from “judge the size of the objects” to “judge the size of the *animals*”, that improves performance on the task). The ISC-CI model extends this idea by inferring a context representation given a set of objects that co-occur in that context. This allows the model to rapidly alter its context representation trial-by-trial (e.g., by attending to different semantic features in each trial of an inductive inference task). This mechanism provides the model with a considerable degree of flexibility, as it can potentially use a different context representation for each set of objects under consideration. It may also relate closely to, and thus provide a useful reference for understanding, the role of attention heads and “in context learning” in the transformer architectures underlying current LLMs — a topic to which return further on.

Abstraction and Relational Representations

Representational Averaging and the Relational Bottleneck

While the ISC-CI model builds on the idea that context representations used for control are shaped by the same statistical learning mechanisms that underlie the formation of semantic representations, the mechanisms it uses for context inference are different, and may provide useful insights into the relationship between statistical learning, inference and control. Specifically, the ISC-CI model uses a simple form of recurrence that takes the mean of the context-independent representations of the objects that co-occur in a given context. This averaging mechanism enforces a simple form of abstraction by obscuring the identity of individual objects, which highlights the relationships between the objects rather than the details of each object, minimizing the differences between the representations while emphasizing their similarities. Given the nature of the representations learned, this appears to be sufficient to account for the patterns of inference and reasoning modeled in this article.

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

The ISC-CI model implements a weaker form of relational bottleneck, that retains semantic feature information, but shapes this by the extent to which features are *shared* by the objects under consideration. This allows it to capture *graded* relations (e.g., degrees of similarity) simultaneously along *multiple* dimensions. For example, the representation in the context layer for the set {robin, cardinal} will be similar to that for the set {robin, raven}, insofar as both involve birds; but at the same time it will differ insofar as robins and cardinals share colors that are similar whereas robins and ravens do not. The advantages and disadvantages of this approach are complementary to stronger forms of relational bottleneck, which have been used to capture simpler and more categorical forms of relations (e.g., Kerr et al., 2022; Webb et al., 2021; 2024; 2024b): The relational representations in ISC-CI are better suited to cases in which the precise identity of the objects matters, as in semantic inference, but are less applicable to cases in which abstract relationships matter more than the perceptual or semantic details, as in abstract reasoning tasks (such as Raven's Progressive Matrices; Raven, 2003). Both capabilities are reflected in human cognition, suggesting that it relies on a combination of both types of mechanisms.

Representational Subtraction and Analogical Reasoning

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

For example, if the model receives the objects bear and salmon as input, it could be trained to pass the subtracted representation ($bear - salmon$) through a context layer and predict from this representation that the relationship is *eats*. In contrast, if the model receives the objects frog and tadpole, it could be trained to predict that the relationship is *parent*. This form of relational inference may in turn support analogical reasoning; by processing a support set containing {bear, salmon} and a query set containing {bird}, the model can jointly predict that a) the most likely relation between bear and salmon is *eats* and b) when this relation is applied to

bird the most likely resulting object is worm. This process simulates the analogy “bear is to salmon as bird is to what?” Given such explicit training on object relations, this method might not only be more successful in solving analogy problems than standard word embedding methods (Mikolov et al., 2013) but, critically, be able to explicitly report the dimensions used in doing so, much as people can do.

Relationship to Machine Learning Models of Natural Language Processing

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

Co-Occurrence and Coherent Variation in Language Models

Many computational approaches to understanding semantic structure from natural language build on the idea that the lexical co-occurrences between words is closely tied to the meaning of those words (Wittgenstein, 1953). This approach was first applied successfully by cognitive scientists in the earliest recurrent neural networks (e.g. Elman, 1990; 1991), as well as closely related computational approaches such as latent semantic analysis (LSA; Landauer & Dumais, 1997; Dumais, 2005) and holistic analog to language (HAL; Lund & Burgess, 1996). Later, deep learning approaches such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) scaled this approach to large text corpora, learning vector representations for each word that are used to predict which words co-occur within a fixed window of the target word. Although these models successfully capture some aspects of human semantic structure, they provide relatively poor models of human inductive inference (XX?) and similarity judgments (Iordan et al., 2022; Pereira et al., 2016). One reason for this limitation may be that these models lack the ability to determine which semantic features are

relevant in a given context: Since they only take into account pairwise relationships between words, they always use the same representation for a given word, regardless of the context.

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

Although ISC-CI shares many similarities with LLMs based on the transformer architecture, these differ both in the particulars of the training datasets and their architectures. With respect to training, LLMs are trained on large datasets of natural language, from which they learn to represent words based on patterns of co-occurrence between words in articles, books, and posts on the internet. In contrast, the ISC-CI model is trained on a small dataset meant to represent natural contexts, and the model learns to represent *objects* based on patterns of co-occurrence between those objects in natural contexts. This form of training significantly limits the flexibility of the ISC-CI model (e.g., it can only represent the concrete objects on which it was trained rather than a large set of words), but it allows the model to learn using a significantly smaller dataset that more closely matches human experience (e.g., the ISC-CI model experiences on the order 100,000 training episodes, whereas LLMs experience on the order of 1 trillion tokens). Furthermore, the ISC-CI model is trained with the dual objective of predicting which objects in the query set appear in the same context as objects in the support

set as well as the context shared by those objects (e.g., the “bird” context for the support set {robin, sparrow}). This helps the model more rapidly learn how to perform context inference, but it requires a dataset that contains context labels.

The models also differ substantially in their architectures. While both rely on context-based modulation of context-independent representations (i.e., attentional control; Cohen et al., 1990; Vaswani, 2017), the ISC-CI model relies on a simple averaging mechanism for extracting context rather than the more complex and powerful transformer architecture. We implemented the averaging mechanism as the simplest possible method for extracting context from co-occurring objects. This design choice allowed the model to learn significantly faster; in a pilot study, we found that training a transformer on the same objective as the ISC-CI model took ten times as many training episodes to converge to the same loss as the ISC-CI model. Integration is also a mechanism that is neurally plausible. Thus, its success suggests that it may capture important features of how the learning of context representations and their use for inference are implemented in the brain. Furthermore, the ISC-CI model develops an explicit representation of the context (in the context layer of the model). This allows it to report contexts explicitly, paralleling human capabilities. This also allows it to learn to represent the relationships among different contexts, and thereby the ability to share knowledge between contexts. For example, if the model uses similar representations for the contexts formed by the sets {robin, sparrow} and {robin, canary}, the model can infer that a newly learned fact about one of those contexts (e.g., that *can-fly* is a useful semantic feature for the {robin, sparrow} context) likely applies to the other context as well (e.g., *can-fly* is also useful for {robin, canary}), greatly speeding learning. In contrast, transformer models represent context implicitly in the weights of the network, which may or may not demonstrate the same cross-context knowledge sharing.

Relationship to Bayesian models of induction

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

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

For instance, observing a robin and a raven both labelled "fep," to determine the category "fep" to which refers, the system will first compute the *likelihood* of sampling a robin and a raven from each of many possible categories—animals, vehicles, black things, insects, birds, etc. While many of these categories will have zero probability of generating the two items (e.g. *vehicles*, *things made of metal, only robins and nothing else),* a subset will have non-zero probability (e.g. *animals*, *living things*, *birds*, *flying things*, *things with legs*, etc); and, among those, some will have relatively high probability (e.g. the category containing small forest birds). Among all possible categories, one is guaranteed to have extremely high probability of

producing the two items—specifically, the category that contains just the observed robin and the observed raven and nothing else.

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

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

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

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

The Bayesian program has tackled both issues with substantive research over several years —for instance, exploring compute-efficient ways of approximating large probability distributions (e..g, Levy et al., 2009), and considering hierarchical Bayesian methods for setting priors across a broader and more flexible variety of representational structures (Kemp & Tenenbuam, 2009). The current work suggests an alternative path toward a similar end. Because it eschews discrete category representations in favor of continuous representational

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

Relationship to classic models of induction

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

These constructs serve a role similar to that of prior probabilities in the Bayesian approach, allowing both proposals to capture something important about human-perceived similarity and

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

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

Conclusion

Cognitive theories of both inferential induction and similarity judgments have long faced a conceptual puzzle: human behavior in tasks designed to probe these processes are often well-explained by proximity within a metric semantic representation space, but a host of classic empirical findings suggest that similarity alone is not sufficient to explain a number of

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

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aguilar, C. M., & Medin, D. L. (1999). Asymmetries of comparison. *Psychonomic Bulletin & Review, 6*(2), 328-337.

Attneave, F. (1950). Dimensions of similarity. *The American journal of psychology*, *63*(4), 516-556.

Bayes, T. (1763). *An essay towards solving a problem in the doctrine of chances*. Philosophical Transactions of the Royal Society of London, 53, 370–418.

Bhatia, S. (2023). Inductive reasoning in minds and machines. *Psychological Review*.

Bhatia, S., & Richie, R. (2024). Transformer networks of human conceptual knowledge. *Psychological review*, *131*(1), 271.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Bowdle, B. F., & Gentner, D. (1997). Informativity and asymmetry in comparisons. *Cognitive Psychology, 34*(3), 244-286.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Callaway, F., Rangel, A., & Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLoS computational biology*, *17*(3), e1008863.

Choi, I., Nisbett, R. E., & Smith, E. E. (1997). Culture, category salience, and inductive reasoning. *Cognition, 65*(1), 15-32.

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.

Covington, P., Adams, J., & Sargin, E. (2016, September). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems* (pp. 191-198).

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179-211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, *7*, 195-225.

Giallanza, T., Campbell, D., Cohen, J. D., & Rogers, T. T. (2024). An integrated model of semantics and control. *Psychological Review*.

Giallanza, T., Campbell, D., & Cohen, J. D. (2024b). Toward the emergence of intelligent control: Episodic generalization and optimization. *Open Mind*, *8*, 688-722.

Gilbert, S. J., & Shallice, T. (2002). Task switching: A PDP model. *Cognitive psychology, 44*(3), 297-337.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357-364.

Han, S. J., Ransom, K. J., Perfors, A., & Kemp, C. (2022). Human-like property induction is a challenge for large language models. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44, No. 44).

Han, S. J., Ransom, K. J., Perfors, A., & Kemp, C. (2024). Inductive reasoning in humans and large language models. *Cognitive Systems Research*, *83*, 101155.

Hayes, B. K., Navarro, D. J., Stephens, R. G., Ransom, K., & Dilevski, N. (2019). The diversity effect in inductive reasoning depends on sampling assumptions. *Psychonomic Bulletin & Review, 26*, 1043-1050.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026-1034).

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173-1185.

Hinton, G. (1981). Shape representation in parallel systems. In *Proceedings of the Seventh International Conference on Artificial Intelligence* (pp. 1088-1096).

Hinton, G. E. (1986, August). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society* (Vol. 1, p. 12).

Huang, P. S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2013, October). Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 2333-2338).

Iordan, M. C., 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.

Jang, A. I., Sharma, R., & Drugowitsch, J. (2021). Optimal policy for attention-modulated decisions explains human fixation behavior. *Elife*, *10*, e63436.

Johannesson, M. (2000). Modelling asymmetric similarity with prominence. *British Journal of Mathematical and Statistical Psychology, 53*(1), 121-139.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive science*, *20*(2), 137-194.

Kalanthroff, E., Davelaar, E. J., Henik, A., Goldfarb, L., & Usher, M. (2018). Task conflict and proactive control: A computational theory of the Stroop task. *Psychological review, 125*(1), 59.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263-292.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological review*, *116*(1), 20.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

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.

Levy, R., Reali, F., & Griffiths, T. (2008). Modeling the effects of memory on human online sentence processing with particle filters. *Advances in neural information processing systems*, *21*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

Mcclelland, J. L., & Farah, M. J. (1991). A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology*.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*(3), 419.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109-165). Academic Press.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547-559.

McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. L. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*(3), 517-532.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Musslick, S., Saxe, A., Hoskin, A. N., Sagiv, Y., Reichman, D., Petri, G., & Cohen, J. D. (2020). On the rational boundedness of cognitive control: Shared versus separated representations.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology, 23*(1), 94-140.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. *Formal approaches in categorization*, 18-39.

Ortony, A., Vondruska, R. J., Foss, M. A., & Jones, L. E. (1985). Salience, similes, and the asymmetry of similarity. *Journal of memory and language, 24*(5), 569-594.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological review*, *97*(2), 185.

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*.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). 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.

Plaut, D. C., & Shallice, T. (1993). Perseverative and semantic influences on visual object naming errors in optic aphasia: A connectionist account. *Journal of cognitive neuroscience*, *5*(1), 89-117.

Raven, J. (2003). Raven progressive matrices. In *Handbook of nonverbal assessment* (pp. 223-237). Boston, MA: Springer US.

Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, *741*(659-663).

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of verbal learning and verbal behavior*, *14*(6), 665-681.

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.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, *104*(3), 192.

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(1), 1-28.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In Attention and performance XIV (silver jubilee volume) synergies in experimental psychology, artificial intelligence, and cognitive neuroscience (pp. 3-30).

Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers, 36*, 506-515.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature neuroscience*, *16*(4), 486-492.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, *96*(4), 523.

Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, *39*, 373-421.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317-1323.

Sloman, S. A. (1993). Feature-based induction. *Cognitive psychology*, *25*(2), 231-280.

Slone, L. K., Abney, D. H., Smith, L. B., & Yu, C. (2023). The temporal structure of parent talk to toddlers about objects. *Cognition*, *230*, 105266.

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.

Tang, D., Qin, B., Liu, T., & Yang, Y. (2015, June). User modeling with neural network for review rating prediction. In *Twenty-fourth international joint conference on artificial intelligence*.

Thrun, S., & Pratt, L. (1998). Learning to learn: Introduction and overview. In *Learning to learn* (pp. 3-17). Boston, MA: Springer US.

Tibshirani, R., & Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, *82*(398), 559-567.

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

Tversky, A., & Gati, I. (1978). Studies of similarity. In *Cognition and categorization* (pp. 79-98). Routledge.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, *108*(3), 550.

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological review*, *111*(3), 757.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, …, & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010).

Wang, H., Wang, N., & Yeung, D. Y. (2015, August). Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1235-1244).

Webb, T. W., Frankland, S. M., Altabaa, A., Segert, S., Krishnamurthy, K., Campbell, D., ... & Cohen, J. D. (2024). The relational bottleneck as an inductive bias for efficient abstraction. *Trends in Cognitive Sciences*.

Webb, T., Mondal, S. S., & Cohen, J. D. (2024). Systematic visual reasoning through object-centric relational abstraction. *Advances in Neural Information Processing Systems*, *36*.

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell. (Original work published 1953)

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological review*, *114*(2), 245.