



The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences

Target Article

All authors contributed equally; authorship is in reverse alphabetical order.

Cite this article: Quilty-Dunn J. (2023) The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences* 46, e261: 1–75. doi:10.1017/S0140525X22002849

Target Article Accepted: 21 November 2022
Target Article Manuscript Online: 6 December 2022

Commentaries Accepted: 27 March 2023

Keywords:

animal cognition; automaticity; cognitive architecture; deep learning; dual-process theories; implicit attitudes; infant cognition; language-of-thought; object files; visual cognition

What is Open Peer Commentary? What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 22) and an Authors' Response (p. 67). See [bbsonline.org](https://www.cambridge.org/bbs/online) for more information.

Jake Quilty-Dunn^a , Nicolas Porot^b and Eric Mandelbaum^c

^aDepartment of Philosophy and Philosophy-Neuroscience-Psychology Program, Washington University in St. Louis, St. Louis, MO, USA; ^bAfrica Institute for Research in Economics and Social Sciences, Mohammed VI Polytechnic University, Rabat, Morocco and ^cDepartments of Philosophy and Psychology, The Graduate Center & Baruch College, CUNY, New York, NY, USA
quiltydunn@gmail.com, sites.google.com/site/jakequiltydunn/
nicolasporot@gmail.com, nicolasporot.com
eric.mandelbaum@gmail.com, ericmandelbaum.com

Abstract

Mental representations remain the central posits of psychology after many decades of scrutiny. However, there is no consensus about the representational format(s) of biological cognition. This paper provides a survey of evidence from computational cognitive psychology, perceptual psychology, developmental psychology, comparative psychology, and social psychology, and concludes that one type of format that routinely crops up is the language-of-thought (LoT). We outline six core properties of LoTs: (i) discrete constituents; (ii) role-filler independence; (iii) predicate–argument structure; (iv) logical operators; (v) inferential promiscuity; and (vi) abstract content. These properties cluster together throughout cognitive science. Bayesian computational modeling, compositional features of object perception, complex infant and animal reasoning, and automatic, intuitive cognition in adults all implicate LoT-like structures. Instead of regarding LoT as a relic of the previous century, researchers in cognitive science and philosophy-of-mind must take seriously the explanatory breadth of LoT-based architectures. We grant that the mind may harbor many formats and architectures, including iconic and associative structures as well as deep-neural-network-like architectures. However, as computational/representational approaches to the mind continue to advance, classical compositional symbolic structures – that is, LoTs – only prove more flexible and well-supported over time.

1. Introduction

Mental representations remain the central posits of psychology after many decades of scrutiny. But what are mental representations and what forms do they take in nature? In other words, what is the format of thought? This paper revisits an old answer to this question: The language-of-thought hypothesis (LoTH).

LoTH is liable to evoke memories of the previous century: Foundational discussions about the structure of thought in the 1970s, the rise of connectionism in the 1980s, and debates about systematicity and productivity in the 1990s. Now, well into the twenty-first century, it might seem that LoTH is a relic, like Freud's tripartite cognitive architecture or Skinnerian behaviorism – a topic of historical interest, but no longer at the center of scientific or philosophical inquiry into the mind.

We will argue for the opposite view: In the half century since Fodor's (1975) foundational discussion, the case for the LoTH has only grown stronger over time. The chief aim of this paper is to showcase LoTH's explanatory breadth and power in light of recent developments in cognitive science. Computational cognitive science, comparative and developmental psychology, social psychology, and perceptual psychology have all advanced independently, yet evidence from these disparate fields points to the same overall picture: Contemporary cognitive science presupposes the language-of-thought (LoT).

The theoretical literature on LoTH is massive and extremely important for understanding the hypothesis and its historical roots. Given space constraints, we will have to ignore huge portions of this literature. We aim simply to provide the strongest article-sized empirical case for LoTH. As a result, we're forced to ignore a great deal of empirical evidence in favor of LoTH. Work in syntax, semantics, psycholinguistics, and philosophy-of-mind has often been taken to bolster LoTH (Fodor, 1975, 1987). Although the relevance of linguistics (broadly construed) to LoTH remains strong, we situate largely independent forms of evidence at the center of our case. We focus primarily on areas (e.g., perception, system-1 reasoning, animal cognition) that seem less language-like. If even these apparent problem areas offer

evidence for LoTH, then we should be optimistic about finding evidence for LoTH throughout much of the mind.

In section 2, we specify which systems of representation count as LoTs. Some of the conclusions of this section will be a bit surprising, as the natural inferences one should draw from the standard characterization of LoTH have largely been ignored since the view's inception. Then, in sections 3–6, we marshal evidence for LoTH from across the cognitive sciences. Section 3 reviews recent LoT-based developments in computational cognitive science, section 4 surveys a mass of data from the study of human perception, section 5 considers evidence from developmental and comparative psychology, and section 6 examines evidence from social psychology.

We think that LoTH is indispensable to a computational account of the mind. But the empirical case for the view does not stem from the idea that LoTH is the “only game in town,” which it is not (and never really was). Instead, we contend, LoTH is the *best* game in town. For a wide variety of phenomena, it does the best job of explaining why biological minds work in the peculiar ways they do.

Our defense of LoTH doesn't presuppose a single, large-scale opponent. Broadly speaking, our opponents are reductionists of various stripes, for example, traditional neural reductionists (Bickle, 2003; Churchland, 1981), theorists who reduce LoT-like cognition to natural language (Berwick & Chomsky, 2016; Hinzen & Sheehan, 2013), critics of representationalism (Hutto & Myin, 2013; Schwitzgebel, 2013), associationists (Dickinson, 2012; Papineau, 2003; Rydell & McConnell, 2006), and most prominently in recent years, reductionist deep-learning approaches (LeCun, Bengio, & Hinton, 2015).¹ However, with the exception of deep neural networks (DNNs), we will mostly avoid direct engagement with these views – not because they are not of interest, but because the best counter to reductionism is simply to demonstrate the explanatory successes of LoT-like

representational structures. In the context of system-1 cognition, for example, our primary opponents will be associationists; in the context of perception science, where associationism is less prominent, our foil will be rival iconic/imagistic formats. This focus on multiple corners of cognitive science will demonstrate two rare virtues of LoTH: Its unificatory power across disciplines and its generalizability across content domains.

2. What is a language-of-thought?

Classic defenses of LoTH often equated it with the view that mental representations are *structured* (Fodor, 1987; Fodor & Pylyshyn, 1988). The route from this identification to the “Only Game in Town” argument is simple – mental representations must have some sort of structure for computational explanations to succeed, and if LoTH follows from that simple fact, it's hard to envision viable alternatives. Arguably, this emphasis on structure per se was influenced by the idea that the primary alternatives to LoTH were connectionist models that lacked structured representations altogether (Rumelhart & McClelland, 1986; cf. Smolensky, 1990).

However, we don't assume this dialectic here. The main reason is that we think there are structured (i.e., nonatomic) representations couched in non-LoT-like formats. Iconic representations are perhaps the clearest example. Operations like mental rotation (Shepard & Metzler, 1971) and scanning (Kosslyn, Ball, & Reiser, 1978) are inexplicable without appeal to structured representations, but at least some of those representations seem to have an iconic, rather than LoT-like, representational format (Carey, 2009; Fodor, 2007; Kosslyn, 1980; Quilty-Dunn, 2020b; Toribio, 2011; cf. Pylyshyn, 2002). Other potential formats include analog magnitudes (Carey, 2009; Clarke, 2022; Clarke & Beck, 2021; Mandelbaum, 2013; Meck & Church, 1983), vectors in multidimensional similarity spaces (Gauker, 2011), mental maps (Camp, 2007; Rescorla, 2009; Shea, 2018; Tolman, 1948), mental models (Johnson-Laird, 2006), graphical models (Danks, 2014), semantic pointers (Eliasmith, 2013), pattern-separated representations (Yassa & Stark, 2011; cf. Quiroga, 2020), neural representations at various scales (Barack & Krakauer, 2021), and much more. We're happy to let a thousand representational formats bloom.

We take LoTH to describe a representational format with six distinctive properties beyond merely having structure. Many, perhaps all, of these properties are not necessary for a representational scheme to count as an LoT, and some may be shared with other formats. We regard these properties as (somewhat) independent axes on which a format can be assessed for how LoT-like it is. If LoT is a natural kind, then these properties should cluster together homeostatically – that is, if some properties are instantiated, it raises the probability that others are as well (Boyd, 1999). These six features each expand the expressive power of abstract, domain-general cognition, making it advantageous for them to evolve as a cluster. We also suspect there might be distinct LoTs with only partially overlapping properties, perhaps arising in different species or different systems within the same mind. The properties adumbrated here don't necessarily exhaust the characterization of LoTH. The crux of the paper includes several sections devoted to empirical evidence, and a fuller picture of LoTH will emerge throughout.

Before moving to the list of core LoT properties, some caveats about how our approach differs from classic defenses of LoTH. First, although LoTH is sometimes understood as the hypothesis

JAKE QUILTY-DUNN is assistant professor in the Philosophy department and Philosophy-Neuroscience-Psychology program at Washington University in St. Louis. He has published articles in philosophy of mind, vision science, and cognitive psychology. His work has received the William James Prize from the Society of Philosophy and Psychology, the Richard M. Griffith Award from the Southern Society of Philosophy and Psychology, and the William James Prize from the Association for the Scientific Study of Consciousness.

NICOLAS POROT is assistant professor at Africa Institute for Research in Economics and Social Sciences (AIRESS), Mohammed VI Polytechnic University, Rabat, Morocco. His research interests include the cognitive science of belief, animal minds, moral categorization, and experimental semantics.

ERIC MANDELBAUM is associate professor in the Departments of Philosophy & Psychology at the CUNY Graduate Center and in the Department of Philosophy at Baruch College. Prior to CUNY he taught at the University of Oxford, Yale University, and Harvard University. He has published articles on topics in cognitive and social psychology, language, vision, and philosophy of mind. He is the recipient of several awards including the Robert J. Glushko Prize from the Cognitive Science Society and the Roger N. Shepherd Prize, and has received fellowships and grants from the American Council of Learned Societies, the National Endowment for the Humanities, the Mellon Foundation, and the Templeton Foundation, among others.

that mental representations have the same structure as natural language, this is not our strategy. Although some theorists have posited LoT to explain natural-language processing and even play a constitutive role in the compositional semantics of natural language (Fodor, 1987; Pinker, 1994), our plan is to search for LoTs outside natural-language-guided contexts. We will examine LoT-like structures that are less connected to natural language and thus represent stringent test cases for LoTH: Mid-level vision, nonverbal minds, and system-1 cognition. LoTH as we'll defend it is committed to representational formats that are language-like in some broad respects, but independent characterizations are provided by both the logical character of LoT (i.e., the way it resembles formal languages that may be radically unlike natural language) and the previous theoretical literature on LoTH, which commits to certain distinctive features. As long as one agrees that an important class of mental representations has many or all of these features, there is no need to quibble about the analogy to natural language.

Second, we will avoid direct discussion of two features of thought that have dominated earlier discussions, namely, systematicity and productivity (Fodor & Pylyshyn, 1988). We agree with the widespread view that any format worth calling an LoT must not only have structure, but it must be compositional: It must include complex representations that are a function of simple elements plus their mode of combination (cf. Szabo, 2011). But as Camp (2007) and others argue, this feature is arguably present in various representational forms, including maps, and thus is not sufficient for ensuring an LoT. Compositionality that is fully systematic and productive is very good evidence for LoT-like architectures, but we want to leave open whether some of the LoT-like structures we'll explore are fully systematic and productive. As a historical note, this caveat is in keeping with earlier discussions, in which systematicity and productivity were each considered "a contingent feature of thought" (Fodor, 1987, p. 152) that evidences LoTH rather than a constitutive requirement. This caveat also dovetails with the previous one about relaxing the analogy with natural language – while, for example, recursive productivity might be a key feature of natural language (Chomsky, 2017), we allow that some LoT-based systems may fail to be recursive. Finally, although we believe systematicity and productivity were good arguments for LoTH, the nature of these cognitive features and their presence in biological minds, including nonverbal ones, is well-trodden ground (Camp, 2009; Carruthers, 2009). Because our goal is to point in new directions for LoTH, we will invoke systematicity and productivity sparingly, mostly keeping instead to the six core properties listed below. These properties are intended to capture the spirit of earlier presentations of LoTH – a combinatorial, symbolic representational format that facilitates logical, structure-sensitive operations (Fodor & Pylyshyn, 1988) – while framing an updated discussion more closely tied to contemporary experimental research.

Property 1: Discrete constituents. Typical iconic representations holistically encode features and individuals (Fodor, 2007; Hummel, 2013; Kosslyn, Thompson, & Ganis, 2006), while LoT representations comprise distinct constituents corresponding to individuals and their separable features. In a sentence like "That is a pink square object," the predicate "square" can be deleted without any other constituents being deleted. In an iconic representation of a pink square, the relationship between the individual, its color, and its shape is more intertwined. "Pink square" can be the output of a merge operation (Chomsky, 1995) while

the part of the icon that represents pink and the part that represents square are one and the same.

Property 2: Role-filler independence. LoT architectures have a distinctive syntax: They combine constituents in a way that maintains independence between syntactic roles and the constituents that fill them (Frankland & Greene, 2020; Hummel, 2011; Martin & Doumas, 2020). The role *agent* is present in "John loves Mary" and "Mary loves John." The identity of the role is independent of what fills it ("Mary," "John"). Likewise, each constituent maintains its identity independent of its current role ("John" can be agent or patient). Role-filler independence captures the rule-based syntactic characteristics of LoT-like compositionality: The syntactic structure is typed independently of its particular constituents, and the constituents are typed independently of how they happen to compose on a particular occasion. In map-like representations, for example, changing the spatial position of a marker changes not only the putative "predicate" (e.g., *tree*) but also the spatial content of the marker (e.g., its position relative to other map elements); thus maps fail to exhibit full role-filler independence (Kulvicki, 2015). Similarly, connectionist models that bind contents through tensor products (Eliasmith, 2013; Palangi, Smolensky, He, & Deng, 2018; Smolensky, 1990) can simulate compositionality, but fail to preserve identity of the original representational elements; thus they sacrifice role-filler independence, and with it classical compositionality (Eliasmith, 2013, p. 125ff; Hummel, 2011).

Role-filler independence might seem similar to the property of having discrete constituents, but they're not equivalent. One could posit discrete constituents in an unordered set, for example, without positing a role that maintains its identity across multiple fillers. There's also nothing in the positing of discrete constituents per se that precludes the type-identity of those constituents from shifting in various contexts (e.g., GREEN APPLE and GREEN PEN might be complexes of discrete constituents, but the copresence of APPLE vs. PEN might change the identity of GREEN; Travis, 2001).

Property 3: Predicate–argument structure. One distinctively LoT-like mode of combination is *predication*, in which a predicate is applied to an argument to yield a truth-evaluable structure. Simple sentences like "John smokes" and "Mary is tall" are paradigmatic examples. Other representational formats, such as images and maps, are assessable for accuracy, but often (perhaps always) fail to exhibit truth-evaluable predicate–argument structure (Camp, 2018; Kulvicki, 2015; Rescorla, 2009). We'll usually interpret predicate–argument structure as requiring both discrete constituents and role-filler independence, that is, as requiring constituents that function as predicates and arguments but maintain type-identity, and as having predicative syntactic structures that can be operated on independently of the content of nonlogical constituents. Thus this condition is not merely that the system must be capable of expressing propositions like <John smokes> (a condition that can be met by even the simplest neural nets, where <John smokes> can be represented by an unstructured node), but rather that this predicate–argument structure is instantiated in the representational vehicle itself (see, e.g., Fodor, 1987).

Property 4: Logical operators. One hallmark of LoT architectures is the use of logical symbols like NOT, AND, OR, and IF. These operators are discrete constituents that compose into larger structures, a hallmark of LoT-like symbols more generally. Logical operators don't obviously presuppose subsentential LoT-like structure, because one could imagine appending such operators

to otherwise unstructured formats, or to maps (Rescorla, 2009). But they are one piece of an overall LoT-friendly picture, positing discrete constituents that allow for formal-syntactic operations. For example, consider an operation that runs from A-OR-B and NOT-A to B; even if A and B are atomic symbols or maps, their un-LoT-like properties are irrelevant because the operation is sensitive to the logical structure alone. Finding evidence for explicit, discrete logical operators should therefore increase our credence in LoTH, all else equal. We'll construe logical operators as requiring role-filler independence, in that, for example, negation operators are the same no matter what proposition they negate.

Property 5: Inferential promiscuity. LoT architectures have been useful in characterizing inferential transitions, especially logical inferences (Braine & O'Brien, 1998; Fodor & Pylyshyn, 1988; Quilty-Dunn & Mandelbaum, 2018a; Rips, 1994; cf. Johnson-Laird, 2006). LoT-like representations should not only encode information, but they should be usable for inference in a way that is automatic and independent of natural language.² The automaticity point is important: The theories of logical inference just cited share an appeal to computational processes that transform representations with one logical form into representations with another logical form in accordance with rules that are *built into the architecture* (i.e., merely procedural, not explicitly represented, and thus not amenable to intervention from representational states; Quilty-Dunn & Mandelbaum, 2018b). If these theories are even roughly on the right track, then we should find evidence for logical-form-sensitive computation outside conscious, controlled, natural-language-guided contexts.

Property 6: Abstract conceptual content. LoTH has historically been opposed to concept empiricism, the view that concepts are sensory-based (Barsalou, 1999; Prinz, 2002). It is logically compatible with other core LoT properties that some LoTs might be modality-specific (e.g., different LoT symbol types and/or syntactic rules for each modality). But there is no a priori reason to expect that primitive LoT symbols – unlike, for example, iconic or analog formats – will be limited to a certain range of properties (e.g., sensory properties, the referents of simple concepts for classical empiricists). Thus we should expect (*ceteris paribus*) LoT symbols to represent abstract categories without representing specific details (e.g., a symbol that encodes *bottle* and no particular shape or color). There is therefore a nondemonstrative but bidirectional relationship between LoTs and abstract contents: Many LoTs should be expected to encode abstract content, and abstract content is naturally represented by means of discrete LoT-like symbols.

The hypothesis that these features cluster together generates nontrivial predictions. Once we've isolated a particular representation type, evidence for any two features (e.g., discrete constituents and abstract conceptual content) may look completely different. Nonetheless, LoTH predicts that these sorts of evidence should tend to cooccur. This cooccurrence would be surprising from a theory-neutral point of view, but not from the perspective of LoTH. We will use just this sort of clustering-based approach to mount an abductive, empirical argument for LoTH. We focus on independently identified systems to observe whether these six properties cluster in them: Perception, physical reasoning in infants and animals, and system-1 cognition.

3. LoTs in computational cognitive science

Before we turn to the bulk of our evidence, we first consider the status of LoTH in computational modeling – a topic of pressing

concern as the advance of artificial intelligence (AI) has made LoT appear antiquated to some researchers. LoT-style models naturally grew out of symbolic computation (Fodor, 1975; Schneider, 2011; cf. Field, 1978; Harman, 1973), including “GOF AI” (“Good Old-Fashioned Artificial Intelligence”; Haugeland, 1985). As new computational methods arose that did not presuppose symbolic computation, such as connectionism with its subsymbolic elements, LoT-style architectures grew detractors. With recent successes of subsymbolic deep neural networks (DNNs) (e.g., Google AI's Google Translate, Deep Mind's success with AlphaFold at modeling protein structure and with AlphaZero and MuZero at dominating complex games; Schrittwieser et al., 2020), LoT-like architectures may appear obsolete.

However, LoT has seen a resurgence in a computational framework that has led to breakthroughs within cognitive science: Bayesianism. Because Bayesian models of cognition are based on probabilistic updating, they appear to present alternatives to LoTH, which posits logical inference. However, Bayesian computational psychology naturally complements LoT architectures (Erdogan, Yildirim, & Jacobs, 2015; Goodman & Lassiter, 2015; Goodman, Tenenbaum, Feldman, & Griffiths, 2008b; Goodman, Tenenbaum, & Gerstenberg, 2015; Kemp, 2012; Overlan, Jacobs, & Piantadosi, 2017; Piantadosi & Jacobs, 2016; Piantadosi, Tenenbaum, & Goodman, 2012, 2016; Ullman, Goodman, & Tenenbaum, 2012; Yildirim & Jacobs, 2015). Wedding probabilistic reasoning to symbolic system processing has led to the “probabilistic language-of-thought” (PLoT) (Goodman et al., 2015).

PLoTs share a core set of properties: A set of primitives with basic operations for their combination (such as the lambda calculus, e.g., Church from Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008a). Primitives correspond to atomic concepts, which are recursively combined to form concepts of arbitrary complexity (Fodor, 1998; Quilty-Dunn, 2021). All one must do is define a set of primitives, and a set of rules for combination and the system is capable of constructing a potentially infinite string of well-formed formulae (Chomsky, 1965).

Bayesianism adds probabilistic inference to the traditional LoT machinery. One way of accomplishing this is by having a likelihood function that is noisy (combining this with a preference for simplicity, either because it's explicitly specified as a prior for the system, or because it falls out as a function of other constraints). PLoTs are classical symbolic systems that display all the hallmarks of LoT architectures, such as discrete constituents, role-filler independence, predicate-argument structure, productive and systematic compositionality, and inferential promiscuity. They are also, however, flexible probabilistic computational programs, because all other aspects of symbol processing (e.g., how they are combined, which processes use them, which information gets updated for them, even their basic semantics) can be determined probabilistically.

Versions of the PLoT have made serious progress in a number of specific areas, for example, learning taxonomical hierarchical structures such as kinship (Katz, Goodman, Kersting, Kemp, & Tenenbaum, 2008; Kemp, 2012; Mollica & Piantadosi, 2015), causality (Goodman, Ullman, & Tenenbaum, 2011), number (Piantadosi et al., 2012), analogical reasoning (Chayette & Piantadosi, 2017), theory acquisition (Ullman et al., 2012), programs (Liang, Jordan, & Klein, 2010), mapping sentences to logical form (Zettlemoyer & Collins, 2005), general Boolean concept learning (Goodman et al., 2008a, 2008b), and moral rule learning (Nichols, 2021). The sheer breadth and depth of the Bayesian computational revolution itself provides strong evidence in favor

of the viability of the LoT. Instead of computational psychology showing that the LoT is a stale theory of the past, it shows how robust, flexible, powerful, and necessary the LoT is in order to ground our computational cognitive science in a way that maps onto human data.

The models that best approximate one type of human concept learning (e.g., learning that a *wudsy* is the tallest object that is either blue or green) are ones where a fuller set of classical logical connectives are hard-coded as primitives. For instance, Piantadosi et al. (2016) taught participants Boolean and quantificational concepts, then built different LoT models in a lambda calculus and compared them to the human data (Fig. 1a). They found that the models that least resembled human performance tended to have the least LoT-like structure. Models that lacked built-in connectives and represented only primitive features or similarity to exemplars performed poorly, as did models that merely learned response biases and only represented TRUE and FALSE categorization judgments. LoTs built with a single connective from which all others are constructed (such as NAND or conjunctions of Horn clauses, disjunctions with at most one nonnegated disjunct) fared better, but not as well as LoTs with the full suite of Boolean operators (conjunction, disjunction, negation, conditional, and biconditional), which in turn were outperformed by models supplanted further with built-in (first-order) quantifiers.³ Although *wudsy* is not an ordinary lexical concept, it is a learnable concept for humans and its acquisition is best modeled by an LoT-like architecture. Thus Piantadosi et al.'s findings provide an existence proof for the utility of LoT-like architectures in the acquisition of logically complex, nonlexical concepts.

Bayesian computational psychology provides evidence that we can learn complex concepts by running probabilistic inductions over a distinctive sort of representational system. This system exploits a rich array of discrete constituents (including predicates and logical operators) that compose into predicate–argument structures of the form *A wudsy is an F*; these structures function as inferentially promiscuous hypotheses and incorporate built-in logical operators that obey role-filler independence: In other words, this system is an LoT.⁴

Similar architectures have recently been used to capture representations of geometrical structure (Amalric et al., 2017; Romano et al., 2018; Roumi, Marti, Wang, Amalric, & Dehaene, 2021; Sablé-Meyer, Ellis, Tenenbaum, & Dehaene, 2021a, 2021b). For example, Amalric et al. (2017) gave participants a task: Observe a sequence of dots and guess where the next dot will appear. They developed a “language-of-geometry” (see also Romano et al., 2018) and found that the complexity of descriptions in this language predicted human error patterns. Sablé-Meyer et al. (2021a) modified this language (including, e.g., accommodating curve-tracing). Participants took as long as needed to encode shapes, and then reidentified them after a brief delay (Fig. 1b). Description complexity in Sablé-Meyer et al.'s PLoT (Fig. 1c) predicted the duration of both encoding and reidentification.

Our primary aim in this section is to point out that not all cutting-edge computational cognitive science is opposed to LoTH.⁵ Indeed, some of the most impressive work in this area relies on LoTs to model human cognition. Current DNNs may be less well-equipped to capture these capacities. For example, Sablé-Meyer et al. (2021b) examined performance of French adults, Himba adults (who lacked formal education or lexical items for geometric shapes and didn't grow up in a “carpentered world”), and French kindergartners on an “intruder” task where they had to detect an unusual shape in a crowd of shapes. They

found that performance in humans was most similar to a model where shapes are “mentally encoded as a symbolic list of discrete geometric properties” (Sablé-Meyer et al., 2021b, p. 5). This LoT-like model was contrasted with state-of-the-art deep convolutional neural networks (DCNNs) as well as nonconvolutional DNNs (specifically, variational autoencoders), and the LoT model outperformed the alternatives. Furthermore, PLoTs are capable of encoding domain-general models that underwrite commonsensical reasoning, a well-known limitation of extant DNNs (Peters & Kriegeskorte, 2021; Zhu et al., 2020). Given the expressive flexibility of PLoTs and their ability to model concept acquisition from just a single data point, they exhibit some advantages over DNN architectures (Piantadosi et al., 2016, p. 414; cf. Brown et al., 2020; but see Ye & Durrett, 2022).

To be clear on the dialectic, many theorists are inclined to point to advances in AI as sufficient evidence against the LoTH. PLoTs serve as an existence proof that LoT architectures are useful in computational modeling. Our claim is not that DNNs will never be able to model these data; indeed, because DNNs are universal function approximators, perhaps such a claim is *ipso facto* false. Other learning policies (e.g., meta-learning; Finn, Yu, Zhang, Abbeel, & Levine, 2017) or architectures (e.g., transformers; Vaswani et al., 2017) may turn out to match symbolic models at mimicking acquisition of logically complex concepts and geometrical encoding in humans. We also grant that DNNs are useful for various engineering purposes outside the context of modeling biological competences. Our claim is simply that computational modeling has not left LoT-like symbolic models behind; LoTH remains fruitful in twenty-first-century computational cognitive science.

It is well-understood by contributors to this literature that “the form that [LoT] takes has been modeled in many different ways depending on the problem domain” (Romano et al., 2018, p. 2). The PLoTs used to model geometrical cognition possess discrete constituents that combine recursively to form more complex shapes, exhibiting role-filler independence, and encode abstract geometric “primitives” (Amalric et al., 2017) like symmetry and rotation independently of low-level properties. Other PLoTs used to model (complex) concept acquisition possess all these features plus logical operators and predication. Of course, whether any or all of these PLoTs turn out to be isomorphic to human cognition is still – like most questions in cognitive science – open. The two morals we stress are (a) that many of these models are meant to test concrete representational formats at the algorithmic level, (b) that these models implement LoTs, and (c) that they sometimes match human performance better than competitor models.

4. Perception

LoTH is often framed as a thesis about thought – that is, post-perceptual central cognition. The idea that perception itself might be couched in an LoT is often ignored (cf. Fodor, 1975, Ch. 1; Pylyshyn, 2003). Indeed, characterizations of many anti-LoTH views, for example, concept empiricism, appeal to the hypothesis that conceptual representations have the same format as perceptual representations, implicitly ruling out the possibility of LoT in perception (Machery, 2016; Prinz, 2002).

We propose instead to take it as an empirical question whether LoT-like representations are deployed in perception, and we'll argue that the answer is likely “Yes.” If cognition is largely LoT-like, and perception feeds information to cognition, then

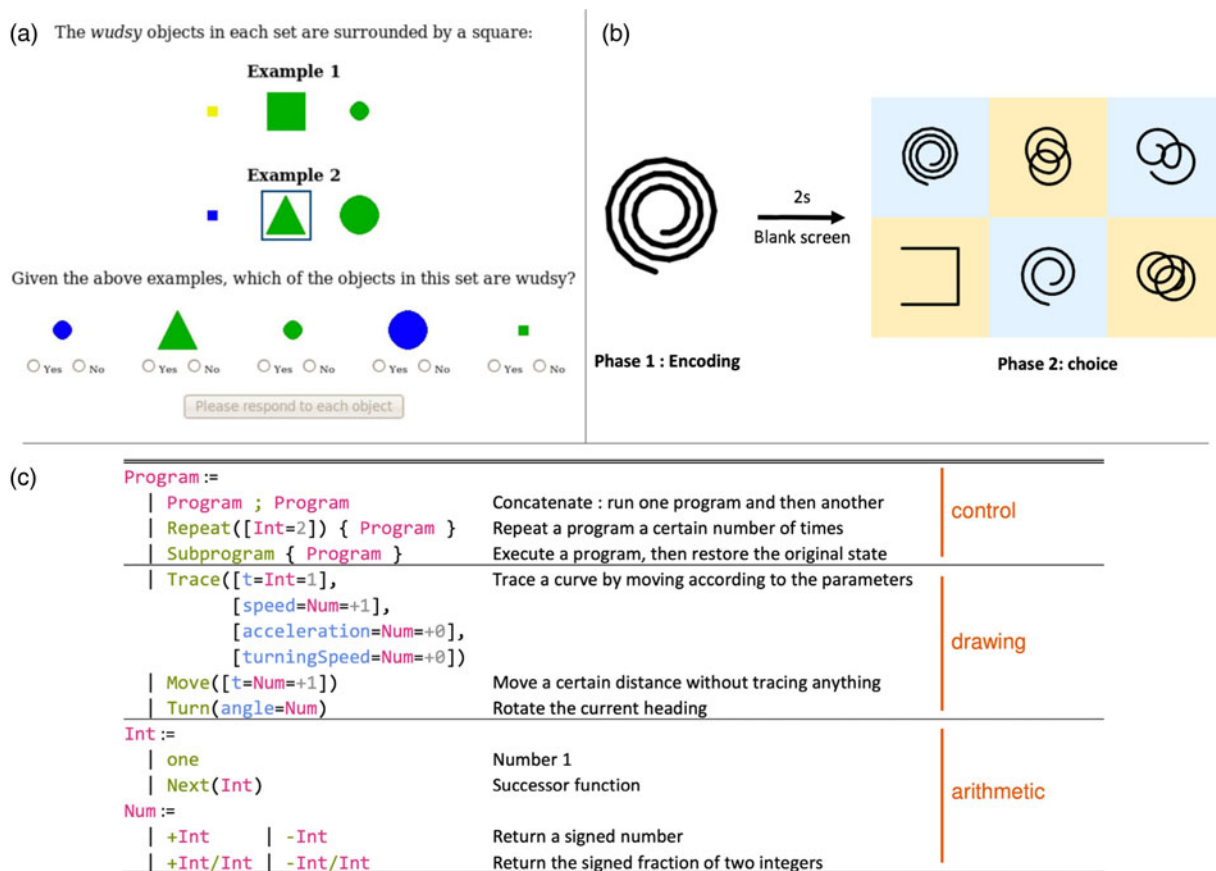


Figure 1. (a) Participants draw inferences about the referent of novel terms like *wudsy* based on examples; reprinted from Piantadosi et al. (2016), Figure 1, with permission from American Psychological Association. (b) Participants encode shapes and reidentify them using minimal description length in a PLoT; reprinted from Sablé-Meyer et al. (2021a), with permission from Mathias Sablé-Meyer. (c) Primitive operations in a geometrical PLoT; reprinted from Sablé-Meyer et al. (2021a), with permission from Mathias Sablé-Meyer.

we should expect at least some elements of perception to be LoT-like, because the two systems need to interface (Cavanagh, 2021; Mandelbaum, 2018; Quilty-Dunn, 2020a). Our case studies include perceptual representations of objects (e.g., object files), relations within objects (e.g., part-whole relations), and relations between objects.

4.1 Object files

Object files are perceptual representations that select individuals, track them across time and space, and store information about them in visual working memory (VWM). This construct is probed via independent, but converging methods, including: Multiple-object tracking (Fig. 2a; Pylyshyn & Storm, 1988), object-based VWM storage (Fig. 2b; Hollingworth & Rasmussen, 2010), physical reasoning, especially in infants (Fig. 2c; Xu & Carey, 1996), and object-specific preview benefits (Fig. 2d; Kahneman, Treisman, & Gibbs, 1992). These methods cluster around a common underlying representation, standardly taken to be a unified representational kind (Carey, 2009; Green & Quilty-Dunn, 2021; Scholl & Leslie, 1999; Smortchkova & Murez, 2020). Object files are extremely well-studied, are generated by encapsulated perceptual processes (Mitroff, Scholl, & Wynn, 2005; Scholl, 2007) that operate prior to and independently of natural-language-guided cognition (Carey, 2009), and are widely believed to have some sort of compositional structure

(minimally, object–property bindings), making them an excellent test case for LoTH.

According to Carey’s (2009) seminal theory of core cognition, object files are amodal but iconic in format (cf. Xu, 2019). Nonetheless, we believe an LoT-based model is better suited to the data than an iconic model (Green & Quilty-Dunn, 2021; Quilty-Dunn, 2020a, 2020c). As far as we know, the possibility of logical operators in object files hasn’t been studied. However, converging evidence suggests that object files have discrete constituents, role-filler independence, predicate–argument structure, and abstract conceptual content. In section 5, we’ll explore the inferential promiscuity of object files in physical reasoning.

4.1.1 First, object files exhibit a decomposition into discrete constituents. Unlike rival models (e.g., iconic models), an LoT-based model of object perception predicts that featural representations should easily break apart from (i) representations of individuals and (ii) other featural representations.

Representations of color and shape frequently come apart from representations of objects without disrupting multiple-object tracking (Fig. 2a) (Bahrami, 2003; Zhou, Luo, Zhou, Zhuo, & Chen, 2010; cf. Pylyshyn, 2007). In VWM, object files dynamically lose featural information like color and orientation independently of one another (Bays, Wu, & Husain, 2011; Fougine & Alvarez, 2011) and VWM resources are depleted independently

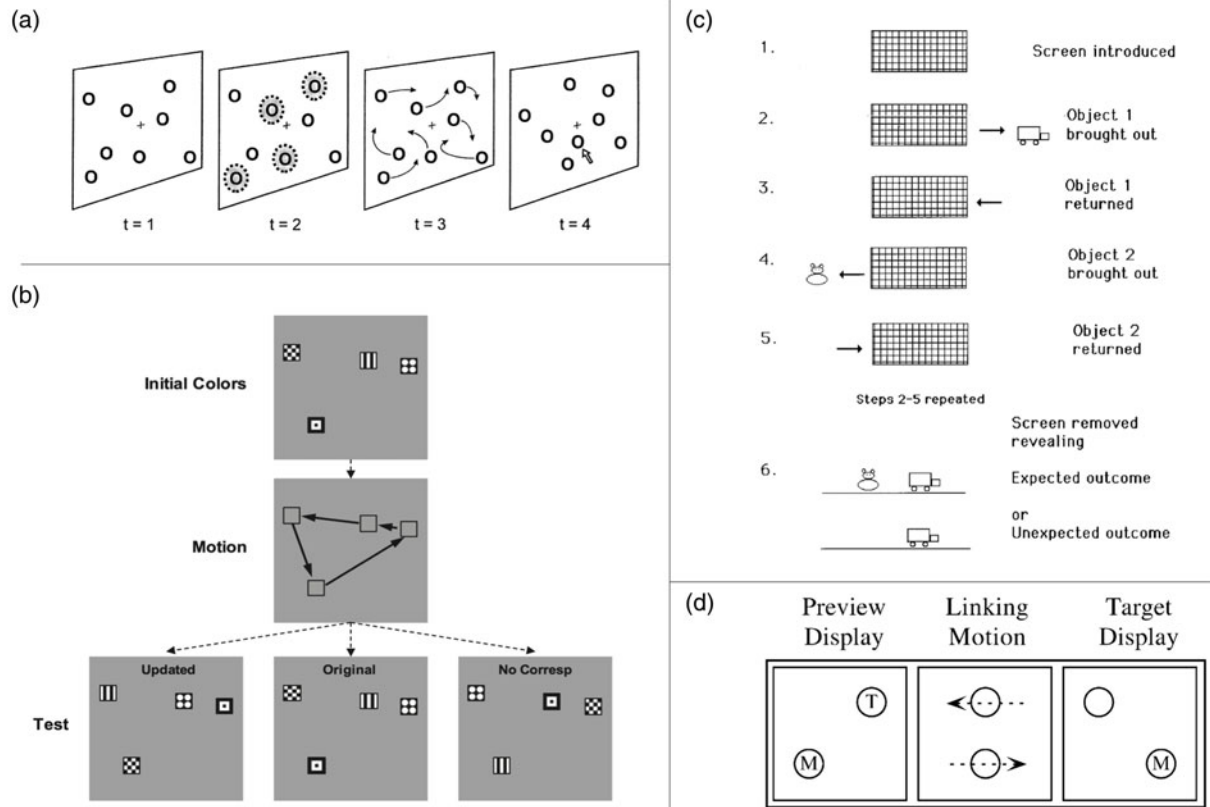


Figure 2. (a) Multiple-object tracking: A subset of visible items (“targets”) is tracked while others (“distractors”) are ignored; reprinted from Pylyshyn (2004), Figure 1, with permission from Taylor & Francis. (b) Object-based VWM storage: A change detection task demonstrates that color is recalled for each object despite location changes, providing just one example piece of evidence that object-based storage in VWM uses object-file representations; reprinted from Hollingworth and Rasmussen (2010), Figure 2, with permission from American Psychological Association. (c) Object-based physical reasoning: Objects pop out from behind an occluder, and preverbal infants rely on spatiotemporal information (and featural and categorical information – see section 5) to keep track of the number of objects, as evidenced by their increased looking time when an unexpected number of items is displayed; reprinted from Xu and Carey (1996), Figure 1, with permission from Elsevier. (d) Object-specific preview benefit: A feature is previewed in each of two visible objects before disappearing, after which the objects move to new locations, and a target feature appears. Subjects show a benefit in reaction time when discriminating the feature if reappears in the same object, illustrating that object-file representations store object properties across spatiotemporal changes; reprinted from Mitroff et al. (2005), Figure 4, with permission from Elsevier.

for color and orientation (Markov, Tiurina, & Utochkin, 2019; Wang, Cao, Theeuwes, Olivers, & Wang, 2017). Similar results hold for real-world stimuli. The state of a book (open or closed) is remembered or forgotten independently of its color or token identity (Brady, Konkle, Alvarez, & Oliva, 2013), and the identity and state of multiple real-world objects are independently swapped in VWM (Markov, Utochkin, & Brady, 2021). These effects are independent of natural-language encoding: They persist when subjects engage in articulatory suppression (Fougnie & Alvarez, 2011; Tikhonenko, Brady, & Utochkin, 2021), and preverbal infants can lose featural information in VWM but maintain a “featureless” pointer-like component of an object file (Kibbe & Leslie, 2011).

In summary, object files in online tracking and VWM appear to break apart freely into discrete constituents, including representations of individuals and separable feature dimensions. This LoT-like format is independent of natural-language capacities.

4.1.2 Second, object files satisfy demanding constraints on predicate–argument structure. One can grant that object files decompose into discrete constituents but deny that these constituents are ordered into a genuinely sentence-like representation. Here we highlight two constraints on genuinely sentence-like predicate–argument representations: Role-filler independence (one of our

six LoT properties) and a grammatical attribution/predication distinction.

Recall that role-filler independence requires that discrete constituents compose into larger structures, but the syntactic structure is typed independently of its particular constituents, and the constituents are typed independently of how they happen to compose on a particular occasion. In a predicate–argument structure in particular, both predicate and argument must maintain type-identity independently of their current bindings – for example, it must be the same JOHN and TALL in TALL(JOHN), TALL(MARY), and SHORT(JOHN).

The clear candidates for predicate-like and argument-like representations in object files are representations of properties and representations of individuals, respectively (cf. Cavanagh, 2021). Representations of individuals must maintain their identity independently of the properties they bind, because tracking performance is successful while properties change (Flombaum, Kundery, Santons, & Scholl, 2004; Flombaum & Scholl, 2006; Zhou et al., 2010) and even while properties are forgotten entirely (Bahrami, 2003; Scholl, Pylyshyn, & Franconeri, unpublished). The computational processes involved in tracking are known as object correspondence processes. Some properties are used to compute object correspondence (e.g., spatiotemporal features and some surface features – see below). However, the fact that the argument-like

representation of the tracked individual can persist while many attributed features are changed/lost entails that the representation maintains independence from the properties to which it is bound.

Likewise, representations of properties maintain their identity independently of the object representations to which they're bound. Some evidence for this is the already-cited fact that they regularly come apart from their respective object representations. However, more striking evidence comes from the way in which featural information is "swapped" between objects. Participants often misremember a feature of one object as bound to another object (Bays, Catalao, & Husain, 2009), including for real-world stimuli (Markov et al., 2021; Utochkin & Brady, 2020). Even during multiple-object tracking, a stored feature of one object (e.g., a previewed numeral) may be swapped with another object if they come too close to each other during tracking (Pylyshyn, 2004). Thus property representations, like individual representations, maintain type-identity across distinct bindings, demonstrating role-filler independence.

The second constraint on predicate–argument structure is a grammatical attribution/predication distinction. In a genuinely sentence-like representation, we can distinguish grammatical positions of predicates. For example:

- (1) That spherical object is red.
- (2) That red object is spherical.

Both attribute spherical shape to the referent of "That," but in (1) the predicate falls within the scope of the noun phrase, whereas in (2) it is in main-predicate position.

One way of capturing this distinction is by appeal to the role of the predicate in grounding the reference of the noun phrase. For example, Perner and Leahy characterize thought in terms of file-like representations (cf. Recanati, 2012), which "capture the predicative structure of language, i.e., the distinction between what one is talking about (the subject, topic, i.e., what the file tracks) and what one says about it (the information about the topic, i.e., the information the file has on it)" (2016, p. 494). Files have "labels" that are captured by (inter alia) determiner phrases like THE RABBIT as well as file contents that include predicates like +FURRY. The attribution of RABBIT in THE RABBIT plays some reference-grounding role, whereas +FURRY is parasitic on the referent of THE RABBIT and merely predicates a property of that referent (see Burge, 2010). In particular, the label-like attributive helps to sustain, and constrain, reference of the file over time.

We can exploit the attribution/predication distinction to see whether the discrete constituents of object files are organized in a genuinely predication-like way, or whether they are merely label-like representations, as in THE RABBIT. The latter format is compatible with an LoT-based model, but part of the virtue of LoTH is that it predicts nontrivial clustering of LoT-like properties. We ought to predict full-blown propositional structures are present in perception as well.

Object files attribute a wide range of properties to their referents, and some of these are used to guide reference to objects. For example, an object file will continue to refer to an object that disappears behind an occluder, but only if it reemerges at a spatiotemporally appropriate location (Scholl & Pylyshyn, 1999). However, although object files attribute other features like color, reference to the object is maintained even if it reemerges a totally different color. Generalizations like this have led some researchers to describe spatiotemporal features as aspects of the object-file "label" while surface features are "stored inside the folder" (Flombaum, Scholl, & Santos, 2009, p. 153). Recent evidence

casts doubt on strict limitations on which properties are part of the "label." although earlier theories took spatiotemporal indices to be uniquely privileged (e.g., Leslie, Xu, Tremoulet, & Scholl, 1998), surface features like color can play an indexing, reference-guiding role in object files, even in ordinary contexts (Hein, Stepper, Hollingworth, & Moore, 2021; Hollingworth & Franconeri, 2009; Moore, Stephens, & Hein, 2010). However, object files routinely store some featural information (e.g., color or orientation) while completely failing to use it to guide reference to objects (e.g., Gordon & Vollmer, 2010; Gordon, Vollmer, & Frankl, 2008; Jiang, 2020; Richard, Luck, & Hollingworth, 2008; see Quilty-Dunn & Green, 2023, for a review).

Object files not only contain discrete constituents, but also the way those constituents are organized satisfies demanding criteria for predicate–argument structure.

4.1.3 Third, object files encode abstract conceptual content. Part of the utility of LoT-like formats is abstracting away from modality-specific information. An LoT allows color and categorical information to be captured in the same representation, as in THAT OBJECT IS A BROWN RABBIT. If object files are LoT-like representations, they not only ought to encode conceptual categories, they ought to do so in a way that abstracts away from sensory details.

The evidence suggests that object files do encode abstract conceptual content. For example, the object-specific preview benefit – a reaction-time benefit in discriminating previously viewed properties of tracked objects (Fig. 2d) – is observed even when the previewed feature is an image of a basic-level category (e.g., APPLE) and the test feature is the corresponding word (e.g., "apple") (Gordon & Irwin, 2000). Similar effects are found for semantic identity of words across fonts (Gordon & Irwin, 1996) or basic-level categories across different exemplars (Pollatsek, Rayner, & Collins, 1984) and across visual and auditory information (Jordan, Clark, & Mitroff, 2010; cf. O'Callaghan, forthcoming). Importantly, these effects do not transfer across associatively related stimuli (e.g., bread–butter), ruling out a reductive associative explanation (Gordon & Irwin, 1996).

Similar effects were recently found in preverbal infants. Kibbe and Leslie (2019) discovered that while infants will not notice whether the first of two serially hidden objects changes its surface features when it reemerges from behind an occluder, they do notice when it changes its category between FACE and BALL. Pomiechowska and Gliga (2021) tested preverbal infants in an EEG change-detection task for familiar categories (e.g., BOTTLE) or unfamiliar categories (e.g., STAPLER). Infants showed an equal response in the negative-central event-related potential (an EEG signature of sustained attention) for across-category and within-category changes for unfamiliar categories, suggesting, unsurprisingly, failure to categorize. But for familiar categories, they showed an increased amplitude only for across-category changes, suggesting that their object files in VWM maintained the conceptual category of the object while visual features decayed.

In adults, VWM seems often to discard specific sensory information in favor of conceptual-category-guided representations (Xu, 2017; 2020; cf. Gayet, Paffen, & Van der Stigchel, 2018; Harrison & Tong, 2009). Participants recall blurry images as less blurry than they really were, suggesting categorical encoding that "goes beyond simply 're-experiencing' images from the past" (Rivera-Aparicio, Yu, & Firestone, 2021, p. 935). Bae, Olkkonen, Allred, and Flombaum (2015) found that object files in online perception and VWM are biased toward the center of color

categories, suggesting that object files store a basic-level color category like RED plus a noisy point estimate within the range of possible red shades. This evidence implicates a category-driven format for object-based VWM representations that abstracts away from low-level visual detail.

Object files encode abstract conceptual content in a way that is not reducible to low-level modality-specific information, just as an LoT-based model predicts.

4.2. Structured relations

We've just argued that perceptual representations of individual objects contain discrete constituents that are organized in a predicate–argument structure and predicate abstract conceptual contents – in other words, they're sentences in the LoT. We'll now describe some LoT-like properties of representations used in the perception of structured relations, both within and between objects.

4.2.1 First, our perceptual systems represent hierarchical part-whole structure. Our perceptual systems don't simply select objects and attribute properties to them. They also break objects down into component parts and represent their part-whole structure. When we perceive a pine tree, we see a branch as part of the tree and a needle as part of the branch, with a sense of the borders between these various parts. Thus the visual system makes use of hierarchical structural descriptions (Fig. 3a; Green, 2019; Hummel, 2013).

The motivation for classic structural-description accounts of object perception was computational: Positing representations of object parts that compose to generate descriptions of part-whole structure allows for successful computational modeling of object perception (Biederman, 1987; Marr & Nishihara, 1978). These models operate just as a classical LoT picture demands, exhibiting systematic and productive compositionality of viewpoint-invariant descriptions of parts (Fig. 3b; Cavanagh, 2021). Structural descriptions “are compositional – forming complex structures by combining simple elements – and thus meaningfully symbolic” (Saiki & Hummel, 1998b, p. 1146).⁶

One of the key assumptions of such models is that object-part boundaries are psychologically real, that is, two points will be treated differently by the visual system when they lie on the same part as opposed to two different parts of the same object. This assumption turns out to be true (Green, 2019). For example, a well-known example of object-based attention is that two stimuli are better discriminated when they lie on the same object than different objects, controlling for distance (Duncan, 1984; Egly, Driver, & Rafal, 1994). The same is true within parts of objects: Participants are quicker to discriminate targets if they lie on the same part than if they cross a part-boundary (Barenholtz & Feldman, 2003). Furthermore, unfamiliar object pairs that share structural descriptions are seen as more similar than object pairs that have a higher degree of overall geometrical similarity but different structural descriptions (Barenholtz & Tarr, 2008).

Role-filler independence emerges directly from structural description models, often explicitly so (Hummel, 2000). Some independent evidence comes from Saiki and Hummel (1998a), who found that shapes of parts and their spatial relations are not represented holistically – in other words, the type-identity of each part is represented independently of its particular role in the structural description and vice versa. Similarity judgments are also guided independently by part shapes and their interrelations, suggesting role-filler independence (Goldstone, Medin, & Gentner, 1991).

We don't deny that the visual system also employs holistic view-based template-like representations (Edelman, 1999; Ullman, 1996) and other formats. Our claims are merely (i) structural descriptions are among the many representations used in visual processing, and (ii) they have an LoT-like format comprising discrete constituents ordered in hierarchical ways that preserve role-filler independence (Fig. 3b).

4.2.2 Second, we perceive structured relations between objects. We don't perceive objects as isolated atoms, as if through a telescope. Instead, we see the glass on the table, the pencils in the cup, and so on.

In a recent review, Hafri and Firestone (2021) survey striking evidence that such relations are recovered rapidly and in abstract form in visual processing (Fig. 3d). For example, the visual system distinguishes containment events (one object disappears inside another) from occlusion events (one disappears behind another) (Strickland & Scholl, 2015). A hallmark of categorical perception is greater discrimination across- than within-category boundaries; participants are better at identifying changes in the position of two circles if the change places the circles in a distinct relation (e.g., CONTAIN(X,Y), TOUCH(X,Y), etc.), suggesting categorically perceived interobject relations (Lovett & Franconeri, 2017). When participants are searching for a particular relation like cup-contains-phone, they are more likely to have a “false-alarm” for target images that instantiate the same relation, like pan-contains-egg, but not book-on-table (Hafri, Bonner, Landau, & Firestone, 2021).

Like structural descriptions, perceptual representations of abstract relations exhibit role-filler independence. Abstract relations apply independently of the relata, and representations of relata persist once the relation is broken – for example, it's the same ON in ON(CAT,COUNTER) and ON(KETTLE,STOVE), and it's the same CAT once the cat leaps off the counter. Hafri et al.'s (2021) finding is especially relevant: The relation CONTAIN(X,Y) governs similarity judgments independently of the relata, about as clear a demonstration of role-filler independence as one could expect to find.

It would be efficient for the visual system to store frequently represented relations. A fascinating recent literature on “scene grammar” (Fig. 3c; Kaiser, Quek, Cichy, & Peelen, 2019; Vö, 2021) details effects of representations of structured relations in visual long-term memory on visual search (Draschkow & Vö, 2017), categorization (Bar, 2004), consciousness (Stein, Kaiser, & Peelen, 2015), and gaze duration (Vö & Henderson, 2009). Relational representations in visual long-term memory (e.g., ON(POT,STOVE) = yes, IN(SPATULA,MICROWAVE) = no) aren't based on associations or statistical summaries over low-level properties. They persist despite changes in position and context (Castelhano & Heaven, 2011), thus abstracting away from overlearned associations. Characteristic scene-grammar effects disappear, however, for upside-down stimuli (Stein et al., 2015), implicating a categorical rather than low-level format. The effects also appear not to rely on summary-statistical information represented outside focal attention (Vö & Henderson, 2009). Despite developing independently of natural language (Öhlschläger & Vö, 2020), structured relations in scene grammar display curious hallmarks of language-like formats. For instance, the P600 ERP increases for syntactic violations in language, and also increases for stimuli that violate visual scene “syntax” (e.g., mouse-on-computer instead of mouse-beside-computer; Vö & Wolfe, 2013). It's standard to

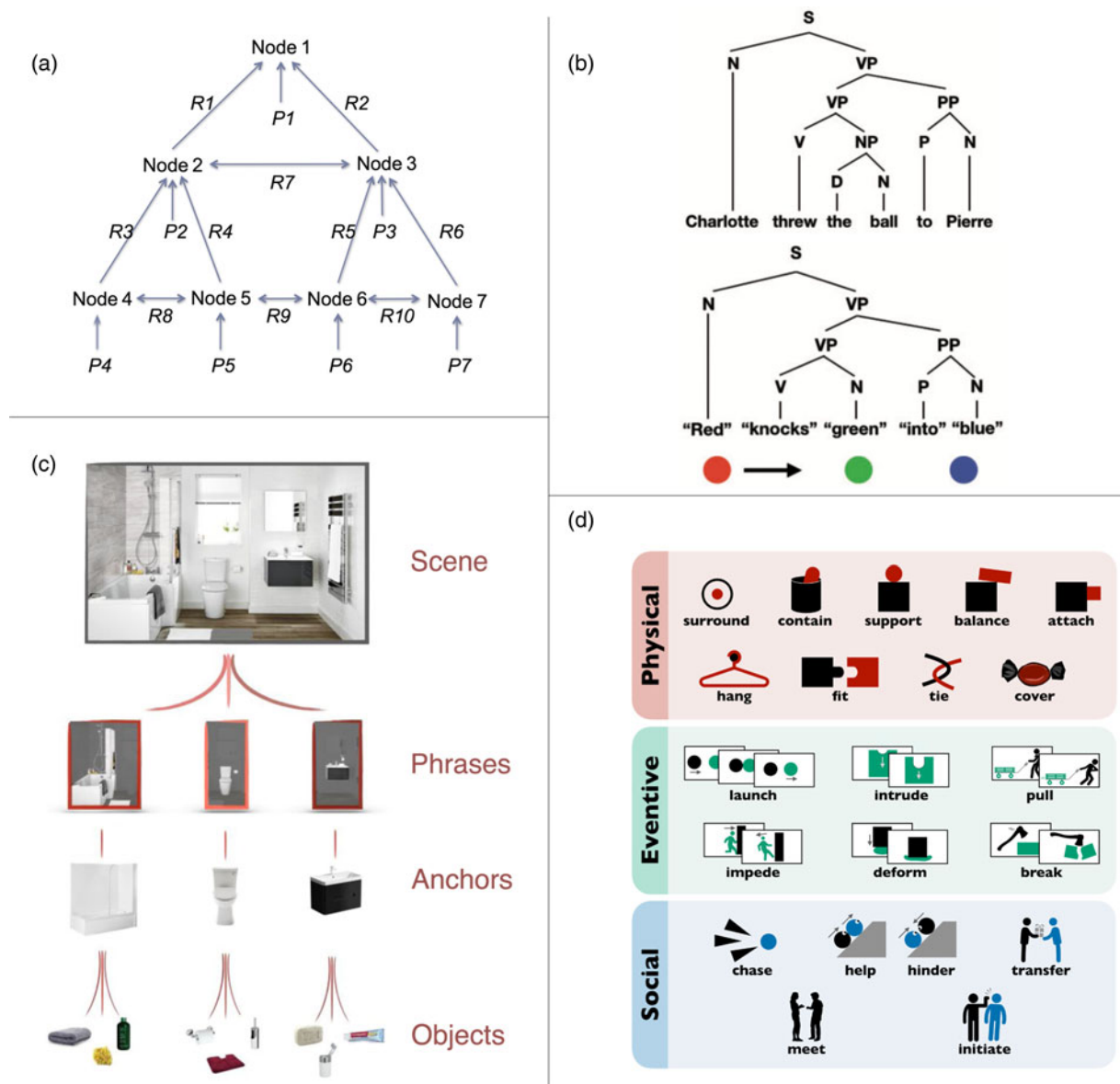


Figure 3. (a) Hierarchical part-whole structural description: Ps = monadic featural properties, horizontal Rs = spatial relations, vertical Rs = mereological relations; reprinted from Green (2019), Figure 9, with permission from Wiley. (b) Structural analogy between tree-like structures in natural-language syntax and tree-like perceptual representations of interobject relations; reprinted from Cavanagh (2021), Figure 3, with permission from Sage under CC BY 4.0, cropped and rearranged. (c) Hierarchical structure in scene grammar: Objects are organized relative to "anchors" (relatively large, immobile elements of environments like showers and trees) in phrase-like structural descriptions of normal relative positions; reprinted from Vö, Bettcher, and Draschkow (2019), Figure 2, with permission from Elsevier. (d) Examples of perceived interobject relations; reprinted from Hafri and Firestone (2021), Figure 2, with permission from Elsevier.

talk of scene grammar as associative, but its relational components satisfy a handful of our LoT hallmarks (e.g., discrete constituents with role-filler independence that encode abstract contents, including categories and relations, and function as arguments in multiplace predicates as in ABOVE(MIRROR, SINK)). Scene grammar is used directly in controlled behavior (e.g., how to arrange a virtual reality scene; Draschkow & Vö, 2017); how broadly it can function in logical inference remains to be explored experimentally.

4.3 Vision and DNNs

In sum, our perceptual capacities to identify and track objects, grasp their characteristic structures, and perceive and store

their relations with one another, appear to rely on LoT-like representations.

A major source of contemporary skepticism about LoTH is the rise of DNNs. Apart from large language models like GPT-3, nowhere are DNNs more visible as models of human cognitive capacities than in visual perception. Given their successes at image classification and apparent similarities to biological vision, one might wonder whether the subsymbolic network structure of DNNs obviates the need to posit LoT-like structures.

The DNNs that have been most touted as models of biological vision are deep convolutional neural networks (DCNNs) trained to classify images (Kriegeskorte, 2015; Yamins & DiCarlo, 2016). After training on large data sets like ImageNet, DCNNs exhibit remarkable levels of performance on image classification.

It is important to evaluate comparisons to human vision not simply in terms of performance, but primarily in terms of underlying competence (Chomsky, 1965). Just as differences in performance need not entail differences in competence (Firestone, 2020), human-like performance on a limited range of tasks need not entail human-like underlying competence. In other words, DCNNs may accomplish image classification while lacking key structural features of human vision, including those relevant to LoTH.

DCNNs have been argued to resemble primate vision in competence as well as performance by appeal to metrics of similarity such as “Representational Similarity Analysis” (Khaligh-Razavi & Kriegeskorte, 2014) and “Brain-Score” (Schrimpf et al., 2018). However, there are shortcomings both to earlier findings of high similarity using these metrics and to the metrics themselves. For example, Xu and Vaziri-Pashkam (2021b) used higher quality fMRI data for their representational similarity analysis and found that, contra Khaligh-Razavi and Kriegeskorte’s earlier findings, high-performing DCNNs (both feedforward and recurrent) show large-scale dissimilarities to human vision. Brain-Score has been criticized for insufficient sensitivity to architectural distinctions (e.g., feedforward vs. recurrent models): “either the Brain-Score metric or the methodology with which a model is evaluated on it fails to distinguish among what we would think of as fundamentally different types of model architectures” (Lonnqvist, Bornet, Doerig, & Herzog, 2021, p. 3). Furthermore, although Schrimpf et al. (2018) found that Brain-Score positively correlates with image classification performance, it fails to capture the crucially hierarchical structure of human vision. Nonaka, Majima, Aoki, and Kamitani (2021) thus developed a “Brain Hierarchy Score” that measures similarities between hierarchical structures, applied it to 29 DNNs, and found a negative correlation between image classification performance and similarity to human vision. This finding provides a striking illustration of how DNNs can excel in performance while veering apart from human competence (see also Fel, Felipe, Linsley, & Serre, 2022).

Our case for LoT in vision is limited to certain domains: Objects, relations between parts and wholes, and relations between objects. It is not a coincidence, in our view, that DNNs that succeed at image classification exhibit little to no competence in these domains. As Peters and Kriegeskorte write about feedforward DCNNs, “the representations in these models remain tethered to the input and lack any concept of an object. They represent things as stuff” (2021, p. 1128).⁷ It is also not clear that DCNNs are capable of representing global shape, let alone the relation between global shape and object parts (Baker & Elder, 2022). Baker, Lu, Erlikhman, and Kellman (2020) trained AlexNet, VGG-19, and ResNet-50 to classify circles and squares, but found that these DCNNs relied only on local contour information; circles made of jagged local edges were classified as squares, and squares made of round local curves were classified as circles. The same models (and several others) also could not distinguish possible from impossible shapes, which requires relating local contour information to global shape (Heinke, Wachman, van Zoest, & Leek, 2021). Failures at processing relations hold not only for DNNs that map images to labels, but also those that map labels to images: Conwell and Ullman (2022) fed the text-guided image-generation model DALL-E 2 a set of interobject relations (including those used by Hafri et al., 2021) and found that it failed reliably to distinguish, for example, “a spoon in a cup” from “a cup on a spoon.”

To be clear, we make no claims about in-principle limitations of DNNs. The machine-learning literature is extremely fast-

moving, and we do not pretend to know what it will look like in even 1 year’s time. Moreover, different DNN architectures might better capture the visual processes discussed here. Although convolutional architectures might privilege local image features, perhaps nonconvolutional architectures like vision transformers (Vaswani et al., 2017) are better suited to avoid these limitations and will supersede DCNNs as models of human vision (Tuli, Dasgupta, Grant, & Griffiths, 2021). Because DCNNs have accumulated enormous publicity despite apparently lacking basic elements of biological vision like global shape and objecthood, future DNN–human comparisons should be approached with caution. Finally, as was noted long ago, neural-network architectures might be able to implement a LoT architecture (Fodor & Pylyshyn, 1988). Indeed, some recent work on DNNs explores implementations of variable binding (Webb, Sinha, & Cohen, 2021; though see Gröndahl & Asokan, 2022; Miller, Naderi, Mullinax, & Phillips, 2022), a classic example of LoT-like symbolic computation (Gallistel & King, 2011; Green & Quilty-Dunn, 2021; Marcus, 2001; Quilty-Dunn, 2021). Our six core LoT properties help specify a cluster of features that such an implementation should aim for.

DNNs are marvels of contemporary engineering. It does not follow that they recapitulate architectural aspects of human vision. We agree with Bowers et al.’s (2022) recent complaint that research on DNNs as models of biological vision is overly focused on performance benchmarks and insufficiently guided by experimental perceptual psychology. Given that DNNs are universal function approximators, and given the vast resources being poured into their development, they will only get closer to human performance over time. But this performance will not reflect core competences of the human visual system unless the relevant models incorporate LoT-like representations of objects and relations.

5. LoTs in nonhuman animals and children

Traditionally, theorists in animal and infant cognition have been reluctant to posit complex cognitive processes, let alone computations over LoT-style representations (e.g., Morgan, 1894; Penn, Holyoak, & Povinelli, 2008; Premack, 2007; cf. Fitch, 2019). However, the state-of-the-art in comparative and developmental psychology is surprisingly congenial to LoTH.

5.1 Abstract content and physical reasoning

Considerable evidence suggests infants use object files to reason about the identity, location, and numerosity of hidden objects (Carey, 2009; Spelke, 1990). However, in a foundational study, Xu and Carey (1996) found that, although 12-month olds who see a duck and then a ball pop out from behind an occluder expect two objects to be present, 10-month olds don’t. This failure might seem to suggest that abstract conceptual content is not usable for physical reasoning in young infants, potentially undermining LoT-based models of infant reasoning (Xu, 2019).

However, 10-month olds do succeed for socially significant categories (Bonatti, Frot, Zangl, & Mehler, 2002; Surian & Caldi, 2010) and objects that are made communicatively salient (Futo, Teglas, Csibra, & Gergely, 2010; Xu, 2019, p. 843). There is also evidence that priming can allow infants to use information in physical reasoning many months earlier than they would otherwise appear to. Lin et al. (2021) made features (e.g., color)

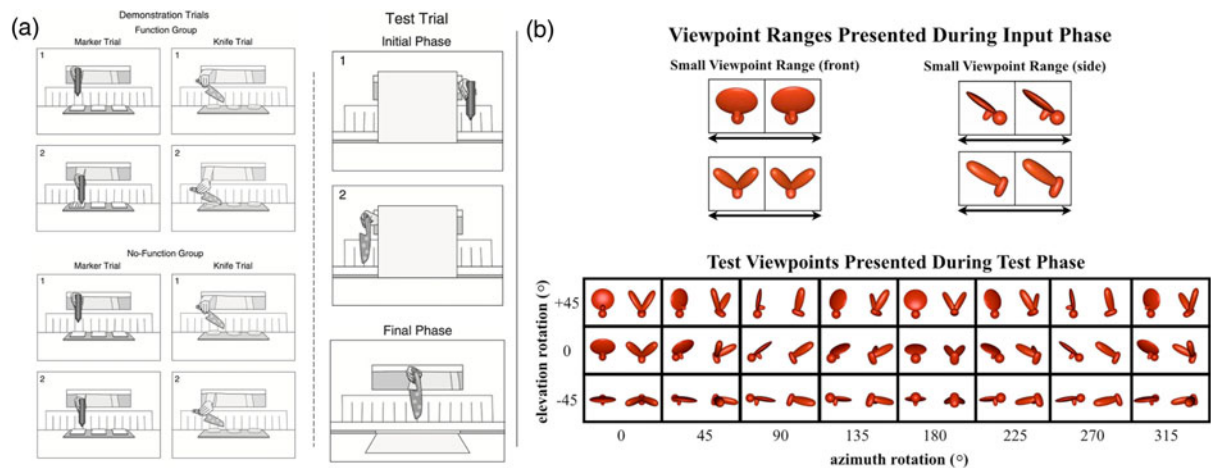


Figure 4. (a) Function demonstrations aid object individuation: In a modification of Xu and Carey's (1996) paradigm, infants first see the characteristic function of an object demonstrated (e.g., a marker drawing, a knife cutting), and this demonstration primes them to use categorical and featural information about the objects to expect two objects in the test trials (i.e., increased looking time when only one object appears); reprinted from Stavans and Baillargeon (2018), Figures 4 and 5, with permission from Wiley. (b) View-invariant information extracted by newborn chicks: Chicks are shown a highly limited set of viewpoints on an object and form an abstract, view-invariant representation; reprinted from Wood and Wood (2020), Figure 1, with permission from Elsevier.

salient by first showing an array of objects that differed along the relevant dimension (e.g., all different colors). This nonverbal priming allowed infants to use information in object files to reason about the individuation of hidden objects 6 months earlier than other methods had detected (e.g., while infants had not shown surprise at a lop-sided object balancing on a ledge until 13 months, Lin et al.'s nonverbal priming of lop-sidedness caused 7-month olds to show the effect).

Infants should therefore be able to use conceptual categories for Xu and Carey's individuation task long before 12 months if the right information is primed first: For example, the relevance of the category's function, a key aspect of artifact concepts (Kelemen & Carey, 2007; cf. Bloom, 1996). Stavans and Baillargeon (2018) demonstrated objects' characteristic functions before hiding (Fig. 4a) and found 4-month olds succeeded at Xu and Carey's individuation task, looking longer when only one object was revealed. These results show two key LoT-like features – abstract content and inferential promiscuity – in extremely young preverbal infants. Thus the earlier failures seem to be explained by performance constraints (Stavans, Lin, Wu, & Baillargeon, 2019).

The use of abstract content in physical reasoning is arguably present throughout the animal kingdom, and is well-studied in primates (e.g., Flombaum et al., 2004) and even some arthropods. Loukola, Perry, Coscos, and Chittka (2017) trained bumblebees through social learning (using a dummy bee) to roll a ball – an unusual behavior for bumblebees in the wild – into the center of a platform for a sucrose reward. When the platform was later rearranged with several balls at various locations that the bees could push into that central area, the bees opted to push balls closest to the center of the platform, even if they differed in color or location from the one they had seen pushed initially. This suggests bumblebees are sensitive to shape in a way that is dissociable from color and location, in contrast to many model-free learning accounts but just as one would expect if shape type is encoded in an LoT. In a similar vein, Solvi, Al-Khudhairy, and Chittka (2020) found that bumblebees could recognize objects under full light that they had previously encountered only in darkness, suggesting they can transfer shape representations stored through touch to a visual task. Bumblebees

therefore appear to represent shape in a way that is dissociable from modality-specific low-level features. These representations figure in practical inferences (thereby displaying inferential promiscuity), and that guides recognition across modalities (thereby displaying abstract content). Furthermore, honeybees trained on a *fewer-than* relation (e.g., $2 < 5$) were able to generalize to cases involving zero items (e.g., $0 < 6$) without any zero-item training, implicating an abstract symbolic representation of *zero* that guides inferential generalization and logico-mathematical reasoning (Howard, Avargues-Weber, Garcia, Greentree, & Dyer, 2018; cf. Vasas & Chittka, 2019; see Weise, Ortiz, & Tibbetts, 2022, for abstract contents of same and different). Similarly, bees' navigational inferences have been used as an argument for a bee LoT because of their computational complexity (Gallistel, 2011).

Much of our discussion in sections 4 and 5.1 has concerned abstract (e.g., amodal or view-invariant) object representations, and one might wonder whether these effects are really because of associations between low-level features acquired gradually during development. One might therefore wonder whether DNNs could therefore provide a better explanation for these effects. However, Wood and Wood (2020) found that newborn chicks showed one-shot learning of abstract object representations (Fig. 4b). Shortly after birth, having been reared in an environment with no movable-object-like stimuli, chicks were shown a virtual three-dimensional (3D)-object rotating either fully 360 degrees, or just 11.25 degrees; later, the chicks successfully recognized the objects from arbitrary viewpoints (equally well under both conditions) and moved toward them. Given the paucity of relevant input, this experiment points away from DNN-based explanations of abstract object representations.

Similarly, Ayzenberg and Lourenco (2021) showed preverbal infants a single view of 60 degrees of an unfamiliar object; using a looking-time measure, they found that the infants formed an abstract, categorical representation, recognizing the object even when viewpoint and salient surface features had drastically changed. The infants' one-shot category learning outperformed DCNNs trained on millions of labeled images. This divergence between DCNN and human performance echoes independent evidence

that DCNNs fail to encode human-like transformation-invariant object representations (Xu & Vaziri-Pashkam, 2021a).

5.2 Logical inference

Proponents of LoTH have long held up its ability to explain logical inference in preverbal children and nonhuman animals as a virtue (Cheney & Seyfarth, 2008; Fodor, 1983; Fodor & Pylyshyn, 1988; Gallistel, 2011; cf. Bermudez, 2003; Camp, 2007, 2009; Gauker, 2011). Recent evidence suggests infants and animals may use logical operators in logical inferences.

Consider the growing body of work on disjunctive syllogistic (DS) reasoning. A standard means of testing for this capacity is Call's (2004) two-cup task. The task involves placing a reward in one of the two cups behind an occluder. Once the cups are brought back into plain view, the participant is shown that one is empty, and can then choose which of the two cups to select from. Typically, researchers are interested in whether the participant selects the unrevealed cup more often than the revealed one, and whether they choose it without inspecting it first. Such behavior is often taken as evidence that the participant can reason through DS, because there's definitely a reward, and one of the two cups is empty, guaranteeing the location of the reward by DS. A surprising number of animals succeed at this task, as well as children as young as two (Call, 2006).

Mody and Carey (2016) argue that there is a confound in such tasks. Participants could rely on a nonlogical strategy involving modal operators: They could form two unrelated beliefs, *MAYBE THERE IS A REWARD IN CUP A* and *MAYBE THERE IS A REWARD IN CUP B*. On this strategy, once shown that cup A is empty, participants simply ignore the possibility that there may be a reward there; left only with the belief that there may be a reward in cup B, they then select cup B. So the authors modified this task, using two rewards and four cups (Fig. 5a). Although children as young as 2.5 succeed at the two-cup task, only 3- and 5-year olds succeed at this four-cup task, with 5-year olds performing best.

Pepperberg, Gray, Cornero, Mody, and Carey (2019) found that an African gray parrot, Griffin, succeeded at a modified version of the four-cup task. Remarkably, Griffin selected the cup that contained reward (a cashew) on nearly every trial (chance, in this case, was 33%), besting human 5-year olds (whose success is surprisingly variable; Gautam, Suddendorf, & Redshaw, 2021). More moderate success at the four-cup task has also been achieved with olive baboons (Fig. 5c; Ferrigno, Huang, & Cantlon, 2021).

A straightforward way of understanding these results is to accept that at least some nonhuman animals are competent with DS. To execute that inference, one needs two sentential connectives, NOT and OR. These must be combined, syntactically, with representations of states of affairs.

The failure of younger kids at Mody and Carey's four-cup task at first looks like bad news for LoTH. However, it might only reflect a failure with using negation, rather than with logical inference more broadly (Feiman, Mody, & Carey, 2022). Moreover, as with Xu's (2019) arguments against LoT-like format in object files, the possibility of performance demands masking an underlying LoT-based competence is plausible. The four-cup task requires kids to track four cups divided into two pairs and two occluded stickers, which is demanding on VWM; indeed, animals that outperform children tend to have superior VWM capacity (Pepperberg et al., 2019, p. 417; cf. Cheng & Kibbe, 2021). As Pepperberg et al. point out, younger children also act more

impulsively than older ones, sometimes ignoring relevant knowledge in demanding tasks. Thus we should look for less demanding tasks before ruling out LoT-like logical inference in children. For example, we could look for independent psychophysical signatures of DS as performed by adults and see whether those signatures are present in children in simpler tasks.

Cesana-Arlotti et al. (2018) showed 12-month olds and adults two objects hidden behind occluders (e.g., a snake and ball); they saw one placed in a cup without knowing which, and finally the unmoved object (e.g., snake) popped out, allowing subjects to infer the identity of the cup-hidden object (ball). When the cup-hidden object was revealed, infants' looking time showed they expected it to be the yet-unseen object (ball). This finding is compatible with nonlogic-based explanations. However, Cesana-Arlotti et al. found that adults performing DS showed an oculomotor signature: During inference, their pupils dilated and eyes darted to the still-hidden object. This same signature was found in the infants, implicating the same underlying computations.

Genuine DS should be domain-general. Cesana-Arlotti, Kovács, and Téglás (2020) used a similar paradigm to test DS in 12-month olds, this time relying on their knowledge of others' preferences. Participants learned an agent's preference among objects (ball vs. car); the nonpreferred object then briefly popped out from behind its occluder, after which the agent reached behind one of the occluders. Twelve-month olds looked longer when the nonpreferred object was reached for. Cesana-Arlotti and Halberda (2022) also found that 2.5-year olds, who fail the four-cup task, nonetheless reason by exclusion across word-learning, social-learning, and explicit negation with a common saccade pattern: They saccade to the to-be-excluded item, return to the target item, and fail to show "redundant" saccades – evidence of low confidence – after target selection. This pattern suggests a domain-general inferential mechanism that delivers high-confidence conclusions, a functional profile one should expect if children perform DS.

Leahy and Carey (2020) provide an alternative, non-DS-based explanation of successful reasoning by exclusion *via* sequentially simulating alternative possibilities. However, chimpanzees, at least, are able to represent distinct possible states of affairs simultaneously. Engelmann et al. (2021) used a modified two-cup task in which the empty cup was not revealed. Chimps could pull ropes for both cups, or pull just one rope for one cup, causing the second cup to fall out of reach. Overwhelmingly they expended extra energy to pull both ropes when the cups were opaque, but pulled just one when the cups were transparent (Fig. 5b).⁸ Pulling two ropes is hedging under uncertainty, suggesting chimps simultaneously represent two locations as possibly reward-laden.

Furthermore, 12-month olds seem to use the same computations adults do to reason by exclusion, as measured by oculomotor signatures (Cesana-Arlotti et al., 2018). It's possible that adults do *both* DS and simulation-based or icon-based reasoning in these tasks. But given independent reasons to think these tasks run on LoT-like object representations in VWM and adults' capacity for DS, and the relative lack of evidence for multiple redundant reasoning processes underlying task performance, our working hypothesis is that infant's oculomotor behavior is evidence for LoT-based DS.

Logical inference without language is a rapidly developing research area, and central contributors to this research such as Carey are skeptical of the "thicker" interpretations of the data we defend. Although we anticipate further plot twists will emerge

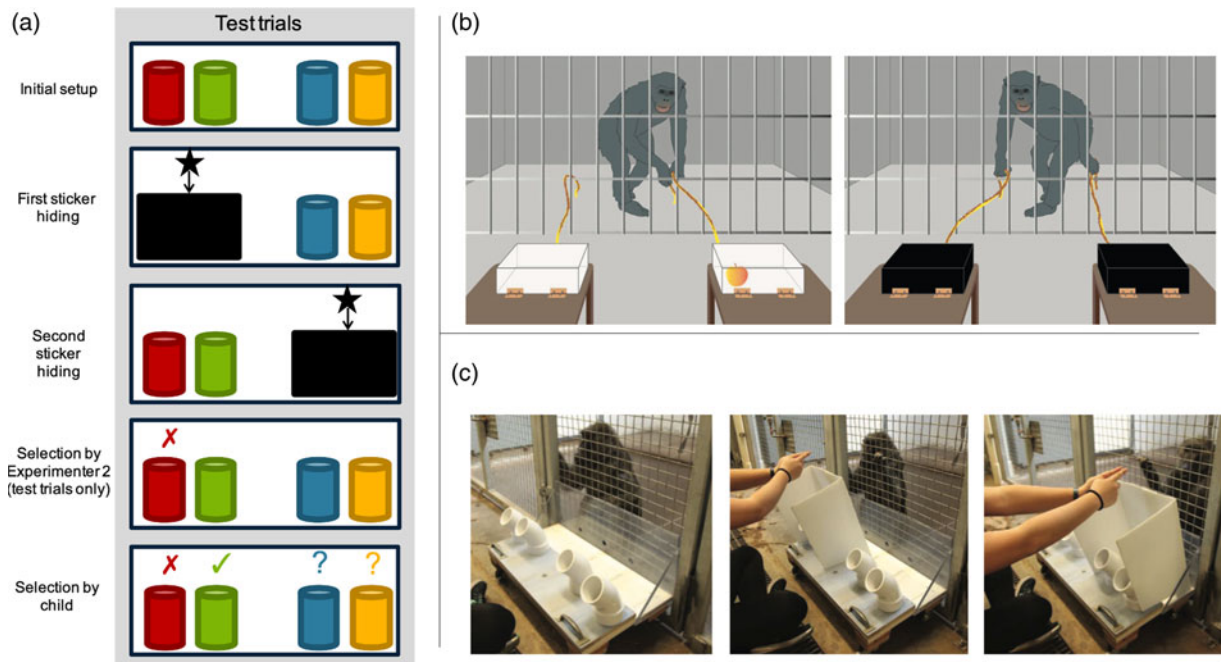


Figure 5. (a) Four-cup task: A reward is placed behind an occluder and into one of the two cups, and again for another reward and pair of cups. Then one cup is shown to be empty, and participants who perform disjunctive syllogism can infer that a reward is certain to be in the other cup in that pair; reprinted from Mody and Carey (2016), Figure 1, with permission from Elsevier. (b) Alternatives in chimps: A reward is placed in one of the two boxes, and chimps pull a string to open the box and reveal the reward. The chimps pull both boxes when they are opaque, suggesting simultaneous representation of two possibilities; reprinted from Engelmann et al. (2021), Figure 1, with permission from Elsevier. (c) Success on four-cup task by baboons, reprinted from Ferrigno et al. (2021), Figure 1, Sage.

in study of infant and nonhuman inference, we take the current state of the literature to favor an LoT-based account of DS in infants and animals and to bear promise for many LoTH-based lines of research in the development of logical operators.

6. LoTs in social psychology: The logic of system 1

One source of opposition to LoTH stems from treatments of attitudes and system-1 processing in social psychology. In traditional dual-process theory, system 1 (“S1”) is governed by shallow heuristic, associative, nonrule-based processing (Evans & Stanovich, 2013; Sloman, 1996). Dual-process theories originate partly from the heuristics-and-biases tradition, where fast responding purportedly demonstrates irrationality (cf. Gigerenzer & Gaissmaier, 2011; Mandelbaum, 2020).

One may doubt the irrationality of S1 processing. As case studies we’ll discuss two paradigms used to investigate characteristically S1 thought: Unconscious reasoning in implicit attitudes in the implicit association test and belief bias cases (though the same morals hold for other paradigms such as base rate inferences and cognitive reflection test; Bago & De Neys, 2017, 2019, 2020; De Neys, Cromheeke, & Osman, 2011; De Neys & Franssens, 2009; De Neys & Glumicic, 2008; De Neys, Rossi, & Houdé, 2013; Gangemi, Bourgeois-Gironde, & Mancini, 2015; Johnson, Tubau, & De Neys, 2016; Pennycook, Trippas, Handley, & Thompson, 2014; Stupple, Ball, Evans, & Kamal-Smith, 2011; Thompson & Johnson, 2014; Thompson, Turner, & Pennycook, 2011).⁹

6.1 Logic, load, and LoT

Failures of syllogistic reasoning are commonplace and well-publicized. In particular, belief biases – cases where people

mistakenly use the truth of a conclusion in judging an argument’s validity, ignoring logical form – are legion (Markovits & Nantel, 1989). Even outside of the belief bias people are forever affirming-the-consequent, denying the antecedent, and confusing validity and truth.

Difficulties in reasoning are *prima facie* problematic for LoTH. The more errors we make in reasoning, the less it seems like we need an inferential apparatus to explain people’s thinking. LoT is tailor-made to explain formal reasoning – that is, reasoning based on the structure, rather than the content, of one’s premises (Fodor & Pylyshyn, 1988; Quilty-Dunn & Mandelbaum, 2018a, 2018b). So, failures in reasoning – traditionally seen as because of heuristic S1 processing – are seen as reasons for believing that S1 is associative rather than LoT-like (see, e.g., Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006; Sloman, 1996). However, a closer look at the data shows evidence for non-associative, LoT-like, logic-sensitive reasoning in S1.

“Conflict problems” are cases where validity and believability conflict, that is, valid syllogisms with unbelievable conclusions or invalid syllogisms with believable conclusions. All other problems (valid/believable; invalid/unbelievable) are “nonconflict.” Some examples:

(Conflict: Valid/Unbelievable)
 P1: All birds fly
 P2: Penguins are birds
 C: Penguins fly

(Conflict: Invalid/Believable)
 P1: All birds fly
 P2: Penguins are birds
 C: Penguins swim

(No Conflict: Valid/Believable)

P1: All birds have feathers

P2: Penguins are birds

C: Penguins have feathers

(No Conflict: Invalid/Unbelievable)

P1: All birds have feathers

P2: Penguins are birds

C: Penguins fly

If S1 is not logic-sensitive, then conflict problems should not hamper believability judgments, because belief bias is driven by nonlogical factors. Yet logic-sensitive judgments occur even when subjects are explicitly instructed to focus on believability, and even under extreme cognitive load. Logical responses thus seem to be generated automatically. People are less confident and slower on conflict problems than nonconflict problems regardless of whether they are judging belief or logic (Handley & Trippas, 2015; Howarth, Handley, & Polito, 2021; Trippas, Thompson, & Handley, 2017). That is, they'll be slower to judge that "Penguins fly" is false if it is a conclusion of a valid argument than a conclusion of an invalid one. Moreover, those who correctly solve syllogism validity questions in conflict problems do so even under intense time pressure and additional memory load, ensuring the shutdown of system-2 processes (Bago & De Neys, 2017). That is, correct responding happens right away; giving participants additional time to think adds little accuracy.

Just as the believability of a conclusion can interfere with validity judgments, so too can the logical form of an argument affect believability judgments. In fact, there is evidence that logical responding is more automatic than belief-based responding; derailing logical responding impedes belief-based responding more than vice versa (Handley, Newstead, & Trippas, 2011; Howarth, Handley, & Walsh, 2016; Trippas et al., 2017). For example, in Trippas et al. (2017), conflict impeded believability judgments more than validity judgments for *modus ponens*. Sensitivity to logical form persists whether subjects are under load or not (and whether asked to evaluate validity or not), showing that the relevant differences are because of S1 processing (Trippas, Handley, Verde, & Morsanyi, 2016). Even when asked to respond randomly, participants still show implicit sensitivity to logical form (Howarth et al., 2021). Automatic logical sensitivity also has very little individual difference between subjects, suggesting it reflects fundamental architectural features of cognition (Ghasemi, Handley, & Howarth, 2021). Logical inferences are also made automatically during reading (Dabkowski & Feiman, 2021; Lea, 1995; Lea, Mulligan, & Walton, 2005). As one would expect if logic was intuitive, subliminally presented premises trigger *modus ponens* inferences (Reverberi, Pischredda, Burigo, & Cherubini, 2012).

Far from undermining LoTH, dual-process architectures vindicate LoTH. They demonstrate abstract logic-based inferential promiscuity outside controlled, conscious cognition using discrete symbols that maintain role-filler independence (e.g., P must be the same symbol in $P \rightarrow Q$).

6.2 The logic of implicit attitudes

Implicit attitudes are typically assumed to be associative. However, Mandelbaum (2016) and De Houwer (2019) documented the effects of "logical interventions" on implicit attitudes, that is, cases where one can change implicit attitudes not by counterconditioning or extinction, as would be expected if they had

associative structure, but instead by merely changing the logically pertinent evidence. Logical (or "propositional") interventions on attitudes are only possible given that we have predicate–argument structure, logical operators, and inferential promiscuity.

Take Kurdi and Dunham (2021). Their basic paradigm consisted of a learning and testing phase. In a learning phase participants saw sentences of the form: "If you see a green circle, you can conclude that Ibbonif is trustworthy; if you see a purple pentagon, you can conclude that Ibbonif is malicious." This design cleverly pits associative versus propositional (i.e., LoT) processes against each other: If the implicit attitude processor is associative then Ibbonif should come out as neutral as Ibbonif is being associated with both positive (trustworthy) and negative (malicious) adjectives. If the processor is sensitive to propositional values however, then the implicit attitude acquired should be dependent on which conditional's antecedent was satisfied (i.e., which shape appears). Participants then moved onto the testing phase which consisted of explicit and implicit attitude testing (via the IAT). Results showed that participant attitudes tracked the logical form of the stimuli during the testing phase. So, using the sample text above, if participants saw a purple pentagon they would conclude that Ibbonif (and the group that he was from, the Niffites, denoted from the suffix on the name) was negatively valenced.

Kurdi and Dunham had ample variations in the paradigm all showing similar LoT-based effects on implicit attitudes. Importantly, LoT-based inferences can be seen *even when the response is normatively inappropriate*, as in an affirming-the-consequent syllogism (study 3). In the learning phase, participants saw sentences such as "If you see a green circle, you can conclude that Ibbonif is malicious"; however, instead of seeing a green circle, they would then see an, for example, orange square. Thus the correct inference to make is that nothing can be inferred from the setup. If implicit attitudes are updated only by an associative processor, then the valence of the predicate in the consequent should dictate the participants' responses. If instead attitudes are sensitive to the logical form of the inventions, then one of the two things should happen: For those subjects who correctly realize that this is an affirming-the-consequent argument they should form no opinion about the person or group in question. However, the subset of people who incorrectly affirm the consequent should make the wrong inference and infer that the consequent accurately describes the person or group in question. Participants were given a control question to see if they were apt to explicitly affirm the consequent. Those who did also changed their implicit attitudes in line with the affirming-the-consequent stimuli they would later see in the experiment; the implicit attitudes of those who rejected the affirming-the-consequent control question, on the contrary, correctly tracked the logical implications of the stimuli by failing to update at all (similar results hold for denying the antecedent). Given a sufficiently creative setup, one can infer logical processes at play even in the *absence* of inference, or during misinference (Quilty-Dunn & Mandelbaum, 2018a).

Similar variations abound. If the associative account were correct then merely giving a major premise that is clearly valenced should set the associative value of the target: Giving participants sentences such as "If you see a purple pentagon, you can conclude that Ibbonif is malicious" should make one associate IBBONIF and negative valence via "malicious." Except that isn't what happens – if subjects are given the conditional premise with no follow-up they withhold forming any valenced implicit attitudes, unlike what associative theory would predict.¹⁰ The concept IBBONIF needs to be linked with the attribute MALICIOUS

in a way that is impervious to associative factors, but sensitive to counterevidence. A predicate–argument structure with MALICIOUS as predicate and IBONIF as argument predicts just this functional profile.

The Kurdi and Dunham is just one of a near-deluge of recent studies showing the efficacy of logical interventions compared to the impotence of associative interventions (Cone & Ferguson, 2015; De Houwer, 2006; Gast & De Houwer, 2013; Mann & Ferguson, 2015, 2017; Mann, Cone, Heggeseth, & Ferguson, 2019; Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016; Van Dessel, Gawronski, Smith, & De Houwer, 2017a; Van Dessel, Mertens, Smith, & De Houwer, 2017b; Van Dessel, Ye, & De Houwer, 2019). Telling participants that they will see a pairing of a group with pictures of pleasant (or unpleasant) things is much more effective at fixing implicit attitudes than repeatedly pairing the group and the pleasant/unpleasant things. One-shot learning trumps 37 associative pairings. Even when associative and one-shot propositional learning are combined, the associative trials add no detectable valence to the implicit attitude formed from the one-shot propositional trial (Kurdi & Banaji, 2017). That is, direct exposure to associative pairings isn't necessary or sufficient for forming or changing implicit attitudes, and its effect on attitudes doesn't compare to a single exposure to a sentence. Even when repeated exposure causes some mental representation of the categories to be formed, just telling participants whether the stimuli are diagnostic modulates learning (e.g., if told the data aren't diagnostic, learning is inhibited, and if told the data are diagnostic, learning is increased). This suggests that the representations acquired are being used as beliefs (Quilty-Dunn & Mandelbaum, 2018b), and updated in a logical, inferentially promiscuous way (Kurdi & Banaji, 2019). The primacy of diagnostic information over repeated exposure is a consistent finding, showing the inadequacies of associative models (e.g., Mann et al., 2019; Mann & Ferguson, 2015, 2017).

In short, implicit attitudes – far from being a problem area for LoT – instead demand evidence-sensitive, inferentially promiscuous predicate–argument structures that incorporate abstract logical operators.

7. Conclusion

More than half a century after the cognitive revolution of the 1950s, mental representations remain the central theoretical posits of psychology. Although our picture of the mind has gotten more and more complex over time, computational operations over structured symbols remain foundational to our explanations of behavior. At least some of these symbols – those involved in certain aspects of probabilistic inference, concept acquisition, S1 cognition, object-based and relational perceptual processing, infant and animal reasoning, and likely elsewhere – are couched in an LoT. That doesn't mean that *all* perceptual and cognitive processing is LoT-symbol manipulation. We believe in other vehicles of thought, including associations (Quilty-Dunn & Mandelbaum, 2020), icons (Quilty-Dunn, 2020b), and much more. Our claim is somewhat modest: Many representational formats across many cognitive systems are LoTs.

We don't deny the successes of DCNNs; perhaps they accurately model some aspects of biological cognition (Buckner, 2019; Shea, 2023). It remains open that DNNs might mimic the performance of biological perception and cognition across a wide variety of domains and tasks by *implementing* core features of LoTs (cf. Zhu et al., 2020). We agree with a recent review of DCNNs that

a “key question for current research is how structured representations and computations may be acquired through experience and implemented in biologically plausible neural networks” (Peters & Kriegeskorte, 2021, p. 1137). Given the evidence above, matching the *competences* of biological minds will require implementing a class of structured representations that uses discrete constituents to encode abstract contents and organizes them into inferentially promiscuous predicate–argument structures that can incorporate logical operators and exhibit role-filler independence.

There is much more to say about evidence for LoT, including abstract, compositional reasoning in aphasics (Varley, 2014), and potential neural underpinnings for LoT (Frankland & Greene, 2020; Gershman, 2022; Roumi et al., 2021; Wang et al., 2019). LoTs ought to provide “common codes” that interface across diverse systems (Dennett, 1978; Pylyshyn, 1973). Central topics here include LoTs at the interfaces of language (Dunbar & Wellwood, 2016; Harris, 2022; Pietroski, 2018) and action (Mylopoulos, 2021; Shepherd, 2021).

The big picture is that LoTH remains a thriving research program. LoTH allows us to distinguish psychological kinds in a remarkably fine-grained way, offering promising avenues for future research. LoTs might differ across systems within a single mind, or between species (Porot, 2019). Although it's likely, for example, that object tracking and S1 reasoning differ in the representational primitives they employ, we don't know whether or how their compositional principles differ. Similarly, we don't know how representations that guide logical inference in baboons differ from those that bees use in social learning, or that infants use in physical reasoning. Differences in conceptual repertoire or syntactic rules provide dimensions along which to type cognitive systems. Future work can focus on decrypting the specific symbols and transformation rules at work in each case, and how these symbols interface with non-LoT mental media.

One might also find subclusters of LoT-like properties. It may be that, for example, properties encoding logical operators and making abstract logical contents available for inference form a “logic” subcluster, and predicate–argument structure, role-filler independence, and abstract contents form a “predication” subcluster. In that case, LoT *qua* natural kind may be a genus of which these subclusters are species (as an analogy, consider how mental icons may be a genus-level kind with high species-level variation between, e.g., visual images and abstract mental models).

Finally, little is known about the evolutionary emergence of LoT in our ancestors or phylogenetically distant LoT-based minds. Our ignorance leaves open the possibility that, given LoTs' computational utility, very different biological minds converged on them independently. An outstanding research goal is to construct a typology of LoTs within and across species, allowing us to better understand the varieties of expressive power in naturally occurring representational systems (Mandelbaum et al., [under review](#)).

Acknowledgments. For helpful discussion and comments, we thank Zed Adams, Jake Beck, Ned Block, Luca Bonatti, Tyler Brooke-Wilson, Chiara Brozzo, Susan Carey, Wayne Christensen, Sam Clarke, Jonathan Cohen, Carl Craver, Roman Feiman, Chaz Firestone, Carolina Flores, E. J. Green, Dan Harris, Zoe Jenkin, Benedek Kurdi, Kevin Lande, Manolo Martinez, Michael Murez, David Neely, Shaun Nichols, Jonathan Phillips, Henry Schiller, Nick Shea, Joshua Shepherd, Brent Strickland, Victor Verdejo, audiences at the Institute of Philosophy in London, Washington University in St Louis, University of Barcelona, Nantes University, the San Diego Winter Workshop in Philosophy of Perception, Cornell University, and the joint meeting of the SPP/ESPP. We are grateful to BBS's reviewers for comments which greatly improved the paper.

Competing interest. None.

Notes

1. We focus on reductionists because one can grant that, e.g., associative processing and natural-language-guided cognition exist, while also positing an LoT. Our opponents are not theorists who merely posit these mechanisms (as we do), but rather theorists who think all *prima facie* LoT-like cognition reduces to them. See, e.g., Lecun et al.'s argument that the success of DNNs "raises serious doubts about whether understanding a sentence requires anything like the internal symbolic expressions that are manipulated by using inference rules" (2015, p. 441).
2. "Usability for inference" here is independent from structural access constraints, e.g., from modularity.
3. Adding second-order quantifiers did not increase performance, suggesting increasing expressive power per se does not necessarily improve model fit.
4. Bayesian modeling is sometimes pitched as a Marrian "computational-level" rational analysis (Anderson, 1990; Oaksford & Chater, 2009). However, a model that better captures human behavior than competitors provides defeasible evidence that some approximation of the computational elements of the model is realized in human cognitive architecture. This "algorithmic-level" approach to computational modeling fits with recent Bayesian approaches (e.g., Lieder & Griffiths, 2020; Vul, Goodman, Griffiths, & Tenenbaum, 2014). We grant that further evidence is needed to establish the algorithmic-level reality of PLoTs (e.g., behavioral evidence of the sort canvassed in the rest of this paper), but we take their success primarily to push back against the dominance of non-LoT-like architectures such as DNNs. Moreover, the fine-grained behavioral measures used in the "language-of-geometry" literature discussed in the next two paragraphs evince an algorithmic-level interpretation.
5. For more critical discussion of DNNs see Lake, Ullman, Tenenbaum, and Gershman (2017) and Marcus (2018).
6. It's possible that "skeletal" shape representations (Feldman & Singh, 2006; Firestone & Scholl, 2014) exhibit similar LoT-like structure (Green, unpublished).
7. Of course DNNs trained for multiple-object tracking do much better (Burgess et al., 2019; Xu, Zhou, Chen, & Li, 2019), but their similarity to human visual competence is underexplored.
8. Chimpanzees, orangutans, monkeys, and children under four fail to hedge in this way when rewards are dropped in a transparent Y-shaped tube: They place a hand under just one of the arms at the bottom (Lambert & Osvath, 2018; Redshaw & Suddendorf, 2016; Suddendorf, Crimston, & Redshaw, 2017; Suddendorf, Watson, Bogaart, & Redshaw, 2019). It is plausible that participants rely on simulation (Leahy & Carey, 2020) here. Unlike the cups task, the Y-shaped tube task requires anticipating the trajectory of an object that is both plainly visible and already in motion, which might encourage simulation.
9. One reason S1 is so instructive is that its operations occur outside working memory. Cognition that is most plausibly governed by internal rehearsal of natural language or "inner speech" plausibly requires verbal-working-memory resources (Baddeley, 1992; Carruthers, 2018; Marvel & Desmond, 2012). Evidence of LoT-like structures in S1 therefore undermines attempts to reduce LoT-like effects to inner speech.
10. We don't deny that there are associations in S1, just that they suffice to explain the data.

References

- Amalric, M., Wang, L., Pica, P., Figueira, S., Sigman, M., & Dehaene, S. (2017). The language of geometry: Fast comprehension of geometrical primitives and rules in human adults and preschoolers. *PLoS Computational Biology*, *13*(1), e1005273.
- Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.
- Ayzenberg, V., & Lourenco, S. F. (2021). One-shot category learning in human infants. *PsyArXiv*. doi:10.31234/osf.io/acymr
- Baddeley, A. (1992). Working memory. *Science (New York, N.Y.)*, *255*(5044), 556–559.
- Bae, G., Olkkonen, M., Allred, S., & Flombaum, J. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*, 744–763.
- Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109.
- Bago, B., & De Neys, W. (2019). The smart system 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299.
- Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, *26*(1), 1–30.
- Bahrami, B. (2003). Object property encoding and change blindness in multiple object tracking. *Visual Cognition*, *10*(8), 949–963.
- Baker, N., & Elder, J. H. (2022). Deep learning models fail to capture the configural nature of human shape perception. *iScience*, *25*(9), 104913.
- Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2020). Local features and global shape information in object classification by deep convolutional neural networks. *Vision Research*, *172*, 46–61.
- Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*, 617–629.
- Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, *22*(6), 359–371.
- Barenholtz, E., & Feldman, J. (2003). Visual comparisons within and between object parts: Evidence for a single-part superiority effect. *Vision Research*, *43*, 1655–1666.
- Barenholtz, E., & Tarr, M. J. (2008). Visual judgment of similarity across shape transformations: Evidence for a compositional model of articulated objects. *Acta Psychologica*, *128*, 331–338.
- Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–609.
- Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 1–11.
- Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, *49*(6), 1622–1631.
- Bermudez, J. L. (2003). *Thinking without words*. Oxford University Press.
- Berwick, R. C., & Chomsky, N. (2016). *Why only us: Language and evolution*. MIT Press.
- Bickle, J. (2003). *Philosophy and neuroscience: A ruthlessly reductive account*. Kluwer.
- Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.
- Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition*, *60*(1), 1–29.
- Bonatti, L., Frot, E., Zangl, R., & Mehler, J. (2002). The human first hypothesis: Identification of conspecifics and individuation of objects in the young infant. *Cognitive Psychology*, *44*, 388–426.
- Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... Blything, R. (2022). Deep problems with neural network models of human vision. *PsyArXiv*. doi:10.31234/osf.io/5zf4s
- Boyd, R. (1999). Homeostasis, species, and higher taxa. In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 141–185). MIT Press.
- Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2013). Real-world objects are not represented as bound units: Independent forgetting of different object details from visual memory. *Journal of Experimental Psychology: General*, *142*(3), 791–808.
- Braine, M. D. S., & O'Brien, D. P. (Eds.). (1998). *Mental logic*. Erlbaum.
- 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.
- Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, *14*(10), e12625.
- Burge, T. (2010). Steps toward origins of propositional thought. *Disputatio*, *4*(29), 39–67.
- Burgess, C. P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., & Lerchner, A. (2019). MONet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv: 1901.11390*.
- Call, J. (2004). Inferences about the location of food in the great apes. *Journal of Comparative Psychology*, *118*, 232–241.
- Call, J. (2006). Descartes' two errors: Reason and reflection in the great apes. In S. Hurley & M. Nudds (Eds.), *Rational animals?* (pp. 219–234). Oxford University Press.
- Camp, E. (2007). Thinking with maps. *Philosophical Perspectives*, *21*, 145–182.
- Camp, E. (2009). Putting thoughts to work: Concepts, systematicity, and stimulus-independence. *Philosophy and Phenomenological Research*, *78*(2), 275–311.
- Camp, E. (2018). Why maps are not propositional. In A. Grzankowski & M. Montague (Eds.), *Non-propositional intentionality* (pp. 19–45). Oxford University Press.
- Carey, S. (2009). *The origin of concepts*. Oxford University Press.
- Carruthers, P. (2009). Invertebrate concepts confront the generality constraint (and win). In R. Lurz (Ed.), *The philosophy of animal minds* (pp. 89–107). Cambridge University Press.
- Carruthers, P. (2018). The causes and contents of inner speech. In A. Vicente & P. Langland-Hassan (Eds.), *Inner speech: New voices* (pp. 31–52). Oxford University Press.
- Castelhano, M. S., & Heaven, C. (2011). Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. *Psychonomic Bulletin & Review*, *18*, 890–896.
- Cavanagh, P. (2021). The language of vision. *Perception*, *50*(3), 195–215.
- Cesana-Arlotti, N., & Halberda, J. (2022). Domain-general logical inference by 2.5-year-old toddlers. *PsyArXiv*. doi:10.31234/osf.io/qzxbp
- Cesana-Arlotti, N., Kovács, A. M., & Téglás, E. (2020). Infants recruit logic to learn about the social world. *Nature Communications*, *11*(5999).
- Cesana-Arlotti, N., Martín, A., Téglás, A., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science (New York, N.Y.)*, *359*, 1263–1266.

- Cheney, D. L., & Seyfarth, R. M. (2008). *Baboon metaphysics: The evolution of a social mind*. University of Chicago Press.
- Cheng, C., & Kibbe, M. M. (2021). Children's use of reasoning by exclusion to track identities of occluded objects. *Proceedings of the Cognitive Science Society*, 43, 2038–2044.
- Cheyette, S., & Piantadosi, S. (2017). Knowledge transfer in a probabilistic language of thought. *Proceedings of the Cognitive Science Society*, 39, 222–227.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.
- Chomsky, N. (1995). *The minimalist program*. MIT Press.
- Chomsky, N. (2017). Language architecture and its import for evolution. *Neuroscience and Biobehavioral Reviews*, 81, 295–300.
- Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78, 67–90.
- Clarke, S. (2022). Beyond the icon: Core cognition and the bounds of perception. *Mind & Language*, 37(1), 94–113. doi:10.1111/mila.12315
- Clarke, S., & Beck, J. (2021). The number sense represents (rational) numbers. *Behavioral and Brain Sciences*, 44, e178.
- Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, 108(1), 37.
- Conwell, C., & Ullman, T. D. (2022). Testing relational understanding in text-guided image generation. *ArXiv*. doi:10.48550/arXiv.2208.00005
- Dabkowski, M., & Feiman, R. (2021). Evidence of accurate logical reasoning in online sentence comprehension. Poster at the Society for Philosophy and Psychology.
- Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. MIT Press.
- De Houwer, J. (2006). Using the implicit association test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, 37(2), 176–187.
- De Houwer, J. (2019). Moving beyond system 1 and system 2. *Experimental Psychology*, 66(4), 257–265.
- De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, 6(1), e15954.
- De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. *Cognition*, 113(1), 45–61.
- De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, 106(3), 1248–1299.
- De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, 20, 269–273.
- Dennett, D. C. (1978). A cure for the common code?. In D. C. Dennett (Ed.), *Brainstorms* (pp. 99–118). Bradford.
- Dickinson, A. (2012). Associative learning and animal cognition. *Philosophical Transactions of the Royal Society B*, 367, 2733–2742.
- Draschlow, D., & Vö, M. L.-H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports*, 7(16471), 1–12.
- Dunbar, E., & Wellwood, A. (2016). Addressing the “two interface” problem: Comparatives and superlatives. *Glossa: A Journal of General Linguistics*, 1(1), 5.
- Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, 123, 501–517.
- Edelman, S. (1999). *Representation and recognition in vision*. MIT Press.
- Egley, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 123, 161–177.
- Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.
- Engelmann, J., Völter, C. J., O'Madagain, C., Proft, M., Haun, D. B., Rakoczy, H., & Herrmann, E. (2021). Chimpanzees consider alternative possibilities. *Current Biology*, 31, R1–R3.
- Erdogan, G., Yildirim, I., & Jacobs, R. A. (2015). From sensory signals to modality-independent conceptual representations: A probabilistic language of thought approach. *PLoS Computational Biology*, 11(11), e1004610.
- Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3), 223–241.
- Feiman, R., Mody, S., & Carey, S. (2022). The development of reasoning by exclusion in infancy. *Cognitive Psychology*, 135, 101473. doi:10.1016/j.cogpsych.2022.101473
- Fel, T., Felipe, I., Linsley, D., & Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. *arXiv preprint arXiv: 2211.04533*.
- Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences of the United States of America*, 103(47), 18014–18019.
- Ferrigno, S., Huang, Y., & Cantlon, J. F. (2021). Reasoning through the disjunctive syllogism in monkeys. *Psychological Science*, 32(2), 292–300.
- Field, H. H. (1978). Mental representation. *Erkenntnis*, 13(1), 9–61.
- Finn, C., Yu, T., Zhang, T., Abbeel, P., & Levine, S. (2017). One-shot visual imitation learning via meta-learning. In *Conference on robot learning* (pp. 357–368). PMLR.
- Firestone, C. (2020). Performance vs. competence in human-machine comparisons. *Proceedings of the National Academy of Sciences of the United States of America*, 117(43), 26562–26571.
- Firestone, C., & Scholl, B. J. (2014). “Please tap the shape, anywhere you like” shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological Science*, 25(2), 377–386.
- Fitch, W. T. (2019). Animal cognition and the evolution of human language: Why we cannot focus solely on communication. *Philosophical Transactions of the Royal Society B*, 375, 20190046. doi:10.1098/rstb.2019.0046
- Flombaum, J. I., Kundery, S. M., Santons, L. R., & Scholl, B. J. (2004). Dynamic object individuation in rhesus macaques: A study of the tunnel effect. *Psychological Science*, 15(12), 795–800.
- Flombaum, J. I., & Scholl, B. J. (2006). A temporal same-object advantage in the tunnel effect: Facilitated change detection for persisting objects. *Journal of Experimental Psychology: Human Perception and Performance*, 32, 840–853.
- Flombaum, J. I., Scholl, B. J., & Santos, L. R. (2009). Spatiotemporal priority as a fundamental principle of object persistence. In B. M. Hood & L. R. Santos (Eds.), *The origins of object knowledge* (pp. 135–164). Oxford University Press.
- Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard University Press.
- Fodor, J. A. (1983). *The modularity of mind*. MIT Press.
- Fodor, J. A. (1987). *Psychosemantics*. MIT Press.
- Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford University Press.
- Fodor, J. A. (2007). The revenge of the given. In B. McLaughlin & J. Cohen (Eds.), *Contemporary debates in philosophy of mind* (pp. 105–116). Blackwell.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3–71.
- Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of Vision*, 11(12), 1–12.
- Frankland, S. M., & Greene, J. D. (2020). Concepts and compositionality: In search of the brain's language of thought. *Annual Review of Psychology*, 71, 273–303.
- Futo, J., Teglas, E., Csibra, G., & Gergely, G. (2010). Communicative function demonstration induces kind-based artifact representation in preverbal infants. *Cognition*, 117, 1–8.
- Gallistel, C. R. (2011). Prelinguistic thought. *Language Learning and Development*, 7, 253–262.
- Gallistel, C. R., & King, A. P. (2011). *Memory and the computational brain: Why cognitive science will transform neuroscience*. John Wiley & Sons.
- Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning – In search of a phenomenon. *Thinking & Reasoning*, 21(4), 383–396.
- Gast, A., & De Houwer, J. (2013). The influence of extinction and counterconditioning instructions on evaluative conditioning effects. *Learning and Motivation*, 44(4), 312–325.
- Gauker, C. (2011). *Words and images: An essay on the origin of ideas*. Oxford University Press.
- Gautam, S., Suddendorf, T., & Redshaw, J. (2021). When can young children reason about an exclusive disjunction? A follow up to Mody and Carey (2016). *Cognition*, 207, 104507. doi:10.1016/j.cognition.2020.104507
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132(5), 692–731.
- Gayet, S., Paffen, S., & Van der Stigchel, S. (2018). Visual working memory storage recruits sensory processing areas. *Trends in Cognitive Sciences*, 22(3), 189–190.
- Gershman, S. J. (2022). The molecular memory code and synaptic plasticity: A synthesis. *arXiv preprint arXiv: 2209.04923*.
- Ghasemi, O., Handley, S. J., & Howarth, S. (2021). The bright homunculus in our head: Individual differences in intuitive sensitivity to logical validity. *Quarterly Journal of Experimental Psychology*, 17470218211044691.
- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62, 451–482.
- Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, 23, 222–262.
- Goodman, N., Mansinghka, V., Roy, D., Bonawitz, K., & Tenenbaum, J. (2008a). Church: A language for generative models. In D. McAllester & P. Myllymaki (Eds.), *Proceedings of the 24th conference on uncertainty in artificial intelligence, UAI 2008* (pp. 220–229). AUAI Press.
- Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics uncertainty in language and thought. In S. Lappin & C. Fox (Eds.), *The handbook of contemporary semantic theory* (pp. 655–686). Wiley.
- Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. (2008b). A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1), 108–154.
- Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Laurence (Eds.), *Concepts: New directions* (pp. 623–654). MIT Press.
- Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, 118(1), 110–199.
- Gordon, R. D., & Irwin, D. E. (1996). What's in an object file? Evidence from priming studies. *Perception and Psychophysics*, 58(8), 1260–1277.
- Gordon, R. D., & Irwin, D. E. (2000). The role of physical and conceptual properties in preserving object continuity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(1), 136–150.
- Gordon, R. D., & Vollmer, S. D. (2010). Episodic representation of diagnostic and non-diagnostic object color. *Visual Cognition*, 18(5), 728–750.
- Gordon, R. D., Vollmer, S. D., & Frankl, M. L. (2008). Object continuity and the transaccadic representation of form. *Perception and Psychophysics*, 70, 667–679.

- Green, E. J. (2019). On the perception of structure. *Noûs*, 53(3), 564–592.
- Green, E. J., & Quilty-Dunn, J. (2021). What is an object file? *British Journal for the Philosophy of Science*, 72(3), 665–699.
- Green, E. J. (unpublished). A pluralist perspective on shape constancy.
- Gröndahl, T., & Asokan, N. (2022). Do transformers use variable binding? *ArXiv*. doi:10.48550/2203.00162
- Hafri, A., Bonner, M. F., Landau, B., & Firestone, C. (2021). A phone in a basket looks like a knife in a cup: The perception of abstract relations. *PsyArXiv*. doi:10.31234/osf.io/jx4yg
- Hafri, A., & Firestone, C. (2021). The perception of relations. *Trends in Cognitive Sciences*, 25(6), 475–492.
- Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(1), 28–43.
- Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: A new parallel processing model. *Psychology of learning and motivation* (Vol. 62, pp. 33–58). Academic Press.
- Harman, G. (1973). *Thought*. Princeton University Press.
- Harris, D. W. (2022). Semantics without semantic content. *Mind & Language*, 37(3), 304–328.
- Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, 458(7238), 632–635.
- Haugeland, J. (1985). *Artificial intelligence: The very idea*. MIT Press.
- Hein, E., Stepper, M. Y., Hollingworth, A., & Moore, C. M. (2021). Visual working memory content influences correspondence processes. *Journal of Experimental Psychology: Human Perception and Performance*, 47(3), 331–343.
- Heinke, D., Wachman, P., van Zoest, W., & Leek, E. C. (2021). A failure to learn object shape geometry: Implications for convolutional neural networks as plausible models of biological vision. *Vision Research*, 189, 81–92.
- Hinzen, W., & Sheehan, M. (2013). *The philosophy of generative grammar*. Oxford University Press.
- Hollingworth, A., & Franconeri, S. L. (2009). Object correspondence across brief occlusion is established on the basis of both spatiotemporal and surface feature cues. *Cognition*, 113(2), 150–166.
- Hollingworth, A., & Rasmussen, I. P. (2010). Binding objects to locations: The relationship between object files and visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 36(3), 543–564.
- Howard, S. R., Avargues-Weber, A., Garcia, J. E., Greentree, A. D., & Dyer, A. G. (2018). Numerical ordering of zero in honeybees. *Science (New York, N.Y.)*, 360, 1124–1126.
- Howarth, S., Handley, S., & Polito, V. (2021). Uncontrolled logic: Intuitive sensitivity to logical structure in random responding. *Thinking & Reasoning*, 28(1), 1–36. doi:10.1080/13546783.2021.1934119
- Howarth, S., Handley, S. J., & Walsh, C. (2016). The logic-bias effect: The role of effortful processing in the resolution of belief–logic conflict. *Memory & Cognition*, 44(2), 330–349.
- Hummel, J. E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 157–185). Erlbaum.
- Hummel, J. E. (2011). Getting symbols out of a neural architecture. *Connection Science*, 23(2), 109–118.
- Hummel, J. E. (2013). Object recognition. In D. Reisburg (Ed.), *Oxford handbook of cognitive psychology* (pp. 32–46). Oxford University Press.
- Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. MIT Press.
- Jiang, H. (2020). Effects of transient and nontransient changes of surface feature on object correspondence. *Perception*, 49(4), 452–467.
- Johnson, E. D., Tubau, E., & De Neys, W. (2016). The doubting system 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, 164, 56–64.
- Johnson-Laird, P. (2006). *How we reason*. Oxford University Press.
- Jordan, K. E., Clark, K., & Mitroff, S. M. (2010). See an object, hear an object file: Object correspondence transcends sensory modality. *Visual Cognition*, 18(4), 492–503.
- Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, 24(2), 175–219.
- Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in Cognitive Sciences*, 23(8), 672–685.
- Katz, Y., Goodman, N. D., Kersting, K., Kemp, C., & Tenenbaum, J. B. (2008). Modeling semantic cognition as logical dimensionality reduction. *Proceedings of the Cognitive Science Society*, 30, 71–76.
- Kelemen, D., & Carey, S. (2007). The essence of artifacts: Developing the design stance. In E. Margolis & S. Laurence (Eds.), *Creations of the mind: Theories of artifacts and their representation* (pp. 212–230). Oxford University Press.
- Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, 119(4), 685–722.
- Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, 10(11), e1003915.
- Kibbe, M. M., & Leslie, A. M. (2011). What do infants remember when they forget? Location and identity in 6-month-olds' memory for objects. *Psychological Science*, 22(12), 1500–1505.
- Kibbe, M. M., & Leslie, A. M. (2019). Conceptually rich, perceptually sparse: Object representations in 6-month-old infants' working memory. *Psychological Science*, 30(3), 362–375.
- Kosslyn, S. M. (1980). *Image and mind*. Harvard University Press.
- Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: Evidence from studies of imagery scanning. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 47–60.
- Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. Oxford University Press.
- Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1(1), 417–446.
- Kulvicki, J. (2015). Maps, pictures, and predication. *Ergo*, 2(7). doi:10.3998/ergo.12405314.0002.007
- Kurdi, B., & Banaji, M. R. (2017). Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? *Journal of Experimental Psychology: General*, 146(2), 194.
- Kurdi, B., & Banaji, M. R. (2019). Attitude change via repeated evaluative pairings versus evaluative statements: Shared and unique features. *Journal of Personality and Social Psychology*, 116(5), 681.
- Kurdi, B., & Dunham, Y. (2021). Sensitivity of implicit evaluations to accurate and erroneous propositional inferences. *Cognition*, 214, 104792.
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.
- Lambert, M. L., & Osvath, M. (2018). Comparing chimpanzees' preparatory responses to known and unknown future outcomes. *Biology Letters*, 14(9), 20180499. <https://doi.org/10.1098/rsbl.2018.0499>
- Lea, R. B. (1995). On-line evidence for elaborative logical inferences in text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(6), 1469–1482.
- Lea, R. B., Mulligan, E. J., & Walton, J. L. (2005). Accessing distant premise information: How memory feeds reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3), 387–395.
- Leahy, B., & Carey, S. (2020). The acquisition of modal concepts. *Trends in Cognitive Sciences*, 24(1), 65–78.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444.
- Leslie, A. M., Xu, F., Tremoulet, P. D., & Scholl, B. J. (1998). Indexing and the object concept: Developing what and where systems. *Trends in Cognitive Sciences*, 2(1), 10–18.
- Liang, P., Jordan, M., & Klein, D. (2010). *Learning programs: A hierarchical Bayesian approach*. Proceedings of the 27th International Conference on Machine Learning, pp. 639–646.
- Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43, 1–60.
- Lin, Y., Li, J., Gertner, Y., Ng, W., Fisher, C. L., & Baillargeon, R. (2021). How do the object-file and physical-reasoning systems interact? Evidence from priming effects with object arrays or novel labels. *Cognitive Psychology*, 125, 101368. doi:10.1016/j.cogpsych.2020.101368
- Lonnqvist, B., Bornet, A., Doerig, A., & Herzog, M. H. (2021). A comparative biology approach to DNN modeling of vision: A focus on differences, not similarities. *Journal of Vision*, 21(10), 1–10.
- Loukola, O. J., Perry, C. J., Coscos, L., & Chittka, L. (2017). Bumblebees show cognitive flexibility by improving on an observed complex behavior. *Science (New York, N.Y.)*, 355, 833–836.
- Lovett, A., & Franconeri, S. L. (2017). Topological relations between objects are categorically coded. *Psychological Science*, 28(10), 1408–1418.
- Machery, E. (2016). The amodal brain and the offloading hypothesis. *Psychonomic Bulletin & Review*, 23, 1090–1095.
- Mandelbaum, E. (2013). Numerical architecture. *Topics in Cognitive Science*, 5(2), 367–386.
- Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*, 50(3), 629–658.
- Mandelbaum, E. (2018). Seeing and conceptualizing: Modularity and the shallow contents of perception. *Philosophy and Phenomenological Research*, 97(2), 267–283.
- Mandelbaum, E. (2020). Assimilation and control: Belief at the lowest levels. *Philosophical Studies*, 177(2), 441–447.
- Mandelbaum, E., Dunham, Y., Feiman, R., Firestone, C., Green, E. J., Harris, D. W., ... Quilty-Dunn, J. (under review). Problems and mysteries of the many languages of thought.
- Mann, T. C., Cone, J., Heggeseth, B., & Ferguson, M. J. (2019). Updating implicit impressions: New evidence on intentionality and the affect misattribution procedure. *Journal of Personality and Social Psychology*, 116(3), 349–374.
- Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology*, 108(6), 823–849.
- Mann, T. C., & Ferguson, M. J. (2017). Reversing implicit first impressions through reinterpretation after a two-day delay. *Journal of Experimental Social Psychology*, 68, 122–127.
- Marcus, G. F. (2001). *The algebraic mind*. MIT Press.
- Marcus, G. F. (2018). Deep learning: A critical appraisal. *arXiv*. doi:1801.00631

- Markov, Y. A., Tiurina, N. A., & Utochkin, I. S. (2019). Different features are stored independently in visual working memory but mediated by object-based representations. *Acta Psychologica*, 197, 52–63.
- Markov, Y. A., Utochkin, I. S., & Brady, T. F. (2021). Real-world objects are not stored in holistic representations in visual working memory. *Journal of Vision*, 21(3), 1–24.
- Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition*, 17(1), 11–17.
- Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London Series B: Biological Sciences*, 200(1140), 269–294.
- Martin, A. E., & Doumas, L. A. A. (2020). Tensors and compositionality in neural systems. *Philosophical Transactions of the Royal Society B*, 375, 20190306. doi:10.1098/rstb.2019.0306
- Marvel, C. L., & Desmond, J. E. (2012). From storage to manipulation: How the neural correlates of verbal working memory reflect varying demands on inner speech. *Brain and Language*, 120(1), 42–51.
- Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes*, 9(3), 320–334.
- Miller, J., Naderi, S., Mullinax, C., & Phillips, J. L. (2022). Attention is not enough. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44(44), 3147–3153.
- Mitroff, S. R., Scholl, B. J., & Wynn, K. (2005). The relationship between object files and conscious perception. *Cognition*, 96, 67–92.
- Mody, S., & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, 154, 40–48.
- Mollica, F., & Piantadosi, S. (2015). Towards semantically rich and recursive word learning models. *Proceedings of the cognitive science conference* (Vol. 37).
- Moore, C. M., Stephens, T., & Hein, E. (2010). Features, as well as space and time, guide object persistence. *Psychonomic Bulletin & Review*, 17(5), 731–736.
- Morgan, L. C. (1894). *An introduction to comparative psychology*. Walter Scott.
- Mylopoulos, M. (2021). The modularity of the motor system. *Philosophical Explorations*, 24(3), 376–393.
- Nichols, S. (2021). *Rational rules: Towards a theory of moral learning*. Oxford University Press.
- Nonaka, S., Majima, K., Aoki, S. C., & Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, 24, 103013.
- Oaksford, M., & Chater, N. (2009). Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, 32(1), 69–84.
- O’Callaghan, C. (forthcoming). Crossmodal identification. In A. Mroczko-Wąsowicz & R. Grush (Eds.), *Sensory individuals, properties, and perceptual objects*. Oxford University Press.
- Öhlschläger, S., & Vö, M. L.-H. (2020). Development of scene knowledge: Evidence from explicit and implicit scene knowledge measures. *Journal of Experimental Child Psychology*, 194(104782), 1–21.
- Overlan, M. C., Jacobs, R. A., & Piantadosi, S. T. (2017). Learning abstract visual concepts via probabilistic program induction in a language of thought. *Cognition*, 168, 320–334.
- Palangi, H., Smolensky, P., He, X., & Deng, L. (2018). Question-answering with grammatically-interpretable representations. The Thirty-Second AAAI Conference on Artificial Intelligence.
- Papineau, D. (2003). Human minds. *Royal Institute of Philosophy Supplements*, 53, 159–183.
- Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin’s mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31, 109–130.
- Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(2), 544–554.
- Pepperberg, I. M., Gray, S. L., Cornero, F. M., Mody, S., & Carey, S. (2019). Logical reasoning by a grey parrot (*Psittacus erithacus*)? A case study of the disjunctive syllogism. *Behaviour*, 156, 409–445.
- Perner, J., & Leahy, B. (2016). Mental files in development: Dual naming, false belief, identity and intentionality. *Review of Philosophy and Psychology*, 7, 491–508.
- Peters, B., & Kriegeskorte, N. (2021). Capturing the objects of vision with neural networks. *Nature Human Behaviour*, 5(9), 1127–1144.
- Piantadosi, S. T., & Jacobs, R. A. (2016). Four problems solved by the probabilistic language of thought. *Current Directions in Psychological Science*, 25(1), 54–59.
- Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, 123(2), 199–217.
- Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, 123(4), 392–424.
- Pietroski, P. M. (2018). *Conjoining meanings: Semantics without truth values*. Oxford University Press.
- Pinker, S. (1994). *The language instinct*. William Morrow.
- Pollatsek, A., Rayner, K., & Collins, W. E. (1984). Integrating pictorial information across eye movements. *Journal of Experimental Psychology: General*, 113(3), 426–442.
- Pomieczowska, B., & Gliga, T. (2021). Nonverbal category knowledge limits the amount of information encoded in object representations: EEG evidence from 12-month-old infants. *Royal Society Open Science*, 8(200782), 1–17.
- Porot, N. J. (2019). *Some non-human languages of thought* (Doctoral dissertation). CUNY Graduate Center.
- Premack, D. (2007). Human and animal cognition: Continuity and discontinuity. *Proceedings of the National Academy of Sciences of the United States of America*, 104(35), 13861–13867.
- Prinz, J. J. (2002). *Furnishing the mind: Concepts and their perceptual basis*. MIT Press.
- Pylyshyn, Z. W. (1973). What the mind’s eye tells the mind’s brain: A critique of mental imagery. *Psychological Bulletin*, 80(1), 1.
- Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and Brain Sciences*, 25, 157–238.
- Pylyshyn, Z. W. (2003). *Seeing and visualizing: It’s not what you think*. MIT Press.
- Pylyshyn, Z. W. (2004). Some puzzling findings in multiple-object tracking: I. Tracking without keeping track of object identities. *Visual Cognition*, 11(7), 801–822.
- Pylyshyn, Z. W. (2007). *Things and places: How the mind connects with the world*. MIT Press.
- Pylyshyn, Z. W., & Storm, R. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, 3(3), 179–197.
- Quilty-Dunn, J. (2020a). Concepts and predication from perception to cognition. *Philosophical Issues*, 30(1), 273–292.
- Quilty-Dunn, J. (2020b). Is iconic memory iconic? *Philosophy & Phenomenological Research*, 101(3), 660–682.
- Quilty-Dunn, J. (2020c). Perceptual pluralism. *Notus*, 54(4), 807–838.
- Quilty-Dunn, J. (2021). Polysemy and thought: Toward a generative theory of concepts. *Mind & Language*, 36, 158–185.
- Quilty-Dunn, J., & Green, E. J. (2023). Perceptual attribution and perceptual reference. *Philosophy and Phenomenological Research*, 106(2), 273–298.
- Quilty-Dunn, J., & Mandelbaum, E. (2018a). Inferential transitions. *Australasian Journal of Philosophy*, 96(3), 532–547.
- Quilty-Dunn, J., & Mandelbaum, E. (2018b). Against dispositionalism: Belief in cognitive science. *Philosophical Studies*, 175(9), 2353–2372.
- Quilty-Dunn, J., & Mandelbaum, E. (2020). Non-inferential transitions: Imagery and association. In T. Chan & A. Nes (Eds.), *Inference and consciousness* (pp. 151–171). Routledge.
- Quiroga, R. Q. (2020). No pattern separation in the human hippocampus. *Trends in Cognitive Sciences*, 24(12), 994–1007.
- Recanati, F. (2012). *Mental files*. Oxford University Press.
- Redshaw, J., & Suddendorf, T. (2016). Children’s and apes’ preparatory responses to two mutually exclusive possibilities. *Current Biology*, 26, 1758–1762.
- Rescorla, M. (2009). Cognitive maps and the language of thought. *British Journal for the Philosophy of Science*, 60(2), 377–407.
- Reverberi, C., Pischedda, D., Burigo, M., & Cherubini, P. (2012). Deduction without awareness. *Acta Psychologica*, 139(1), 244–253.
- Richard, A. M., Luck, S. J., & Hollingworth, A. (2008). Establishing object correspondence across eye movements: Flexible use of spatiotemporal and surface feature information. *Cognition*, 109(1), 66–88.
- Rips, L. J. (1994). *The psychology of proof*. MIT Press.
- Rivera-Aparicio, J., Yu, Q., & Firestone, C. (2021). Hi-def memories of lo-def scenes. *Psychonomic Bulletin & Review*, 28, 928–936.
- Romano, S., Salles, A., Amalric, M., Dehaene, S., Sigman, M., & Figueira, S. (2018). Bayesian validation of grammar productions for the language of thought. *PLoS ONE*, 13(7), e0200420.
- Roumi, F. A., Marti, S., Wang, L., Amalric, M., & Dehaene, S. (2021). Mental compression of spatial sequences in human working memory using numerical and geometrical primitives. *Neuron*, 109, 2627–2639.
- Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. MIT Press.
- Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, 91(6), 995–1008.
- Sablé-Meyer, M., Ellis, K., Tenenbaum, J., & Dehaene, S. (2021a). A language of thought for the mental representation of geometric shapes. *PsyArXiv*. doi:10.31234/osf.io/28mg4
- Sablé-Meyer, M., Fagot, J., Caparos, S., van Kerkoerle, T., Amalric, M., & Dehaene, S. (2021b). Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity. *Proceedings of the National Academy of Sciences of the United States of America*, 118(16), e2023123118.
- Saiki, J., & Hummel, J. E. (1998a). Connectedness and part-relation integration in shape category learning. *Memory & Cognition*, 26(6), 1138–1156.
- Saiki, J., & Hummel, J. E. (1998b). Connectedness and the integration of parts with relations in shape perception. *Journal of Experimental Psychology: Human Perception and Performance*, 24(1), 227–251.
- Schneider, S. (2011). *The language of thought: A new philosophical direction*. MIT Press.
- Scholl, B. J. (2007). Object persistence in philosophy and psychology. *Mind & Language*, 22(5), 563–591.
- Scholl, B. J., & Leslie, A. (1999). Explaining the infant’s object concept: Beyond the perception/cognition dichotomy. In E. Lepore & Z. W. Pylyshyn (Eds.), *What is cognitive science?* (pp. 26–73). Blackwell.
- Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology*, 38(2), 259–290.

- Scholl, B. J., Pylyshyn, Z. W., & Franconeri, S. L. (unpublished). The relationship between property-encoding and object-based attention: Evidence from multiple object tracking.
- Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*. doi:10.1101/407007
- Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., ... Silver, D. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature* 588(7839), 604–609.
- Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. In N. Nottelman (Ed.), *New essays on belief: Constitution, content, and structure* (pp. 75–99). Palgrave Macmillan.
- Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.
- Shea, N. (2023). Moving beyond content-specific computation in artificial neural networks. *Mind & Language*, 38(1), 156–177. doi:10.1111/mila.12387
- Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science (New York, N.Y.)*, 171, 701–703.
- Shepherd, J. (2021). Intelligent action guidance and the use of mixed representational formats. *Synthese*, 198(17), 4143–4162.
- Slooman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1), 3–22.
- Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46, 159–216.
- Smorthkova, J., & Murez, M. (2020). Representational kinds. In J. Smorthkova, K. Dolega, & T. Schlicht (Eds.), *What are mental representations?* (pp. 213–241). Oxford University Press.
- Solvi, C., Al-Khudhairi, S. G., & Chittka, L. (2020). Bumble bees display cross-modal object recognition between visual and tactile senses. *Science (New York, N.Y.)*, 367, 910–912.
- Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, 14, 29–56.
- Stavans, M., & Baillargeon, R. (2018). Four-month-old infants individuate and track simple tools following functional demonstrations. *Developmental Science*, 21, e12500.
- Stavans, M., Lin, Y., Wu, D., & Baillargeon, R. (2019). Catastrophic individuation failures in infancy: A new model and predictions. *Psychological Review*, 126(2), 196–225.
- Stein, T., Kaiser, D., & Peelen, M. V. (2015). Interobject grouping facilitates visual awareness. *Journal of Vision*, 15(8), 1–11.
- Strickland, B., & Scholl, B. J. (2015). Visual perception involves event-type representations: The case of containment versus occlusion. *Journal of Experimental Psychology: General*, 144(3), 570–580.
- Stupple, E. J., Ball, L. J., Evans, J. S. B., & Kamal-Smith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. *Journal of Cognitive Psychology*, 23(8), 931–941.
- Suddendorf, T., Crimston, J., & Redshaw, J. (2017). Preparatory responses to socially determined, mutually exclusive possibilities in chimpanzees and children. *Biology Letters*, 13(6), 20170170. doi:10.1098/rsbl.2017.0170
- Suddendorf, T., Watson, K., Bogaart, M., & Redshaw, J. (2019). Preparation for certain and uncertain future outcomes in young children and three species of monkey. *Developmental Psychobiology*, 62(2), 191–201.
- Surian, L., & Caldi, S. (2010). Infants' individuation of agents and inert objects. *Developmental Science*, 13(1), 143–150.
- Szabo, Z. (2011). The case for compositionality. In W. Hinzen, E. Machery, & M. Werning (Eds.), *The Oxford handbook on compositionality* (pp. 64–80). Oxford University Press.
- Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, 20(2), 215–244.
- Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, 63(3), 107–140.
- Tikhonenko, P. A., Brady, T. F., & Utochkin, I. S. (2021). Independent storage of real-world object features is visual rather than verbal in nature. *PsyArXiv*. doi:10.31234/osf.io/d9c4h
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4), 189–208.
- Toribio, J. (2011). Compositionality, iconicity, and perceptual nonconceptualism. *Philosophical Psychology*, 24(2), 177–193.
- Travis, C. (2001). *Unshadowed thought*. Harvard University Press.
- Trippas, D., Handley, S. J., Verde, M. F., & Morsanyi, K. (2016). Logic brightens my day: Evidence for implicit sensitivity to logical validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9), 1448–1457.
- Trippas, D., Thompson, V. A., & Handley, S. J. (2017). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory & Cognition*, 45(4), 539–552.
- Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. L. (2021). Are convolutional neural networks or transformers more like human vision? *ArXiv*. doi:2105.07197
- Ullman, S. (1996). *High-level vision*. MIT Press.
- Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, 27(4), 455–480.
- Utochkin, I. S., & Brady, T. F. (2020). Independent storage of different features of real-world objects in long-term memory. *Journal of Experimental Psychology: General*, 149(3), 530–549.
- Van Dessel, P., De Houwer, J., Gast, A., Smith, C. T., & De Schryver, M. (2016). Instructing implicit processes: When instructions to approach or avoid influence implicit but not explicit evaluation. *Journal of Experimental Social Psychology*, 63, 1–9.
- Van Dessel, P., Gawronski, B., Smith, C. T., & De Houwer, J. (2017a). Mechanisms underlying approach-avoidance instruction effects on implicit evaluation: Results of a pre-registered adversarial collaboration. *Journal of Experimental Social Psychology*, 69, 23–32.
- Van Dessel, P., Mertens, G., Smith, C. T., & De Houwer, J. (2017b). The mere exposure instruction effect. *Experimental Psychology*, 64(5), 299–314.
- Van Dessel, P., Ye, Y., & De Houwer, J. (2019). Changing deep-rooted implicit evaluation in the blink of an eye: Negative verbal information shifts automatic liking of Gandhi. *Social Psychological and Personality Science*, 10(2), 266–273.
- Varley, R. (2014). Reason without much language. *Language Sciences*, 46, 232–244.
- Vasas, V., & Chittka, L. (2019). Insect-inspired sequential inspection strategy enables an artificial network of four neurons to estimate numerosity. *iScience*, 11, 85–92.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Vö, M. L.-H. (2021). The meaning and structure of scenes. *Vision Research*, 181, 10–20.
- Vö, M. L.-H., Bettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, 29, 205–210.
- Vö, M. L.-H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, 9(3), 1–15.
- Vö, M. L.-H., & Wolfe, J. M. (2013). Different electrophysiological signatures of semantic and syntactic scene processing. *Psychological Science*, 24(9), 1816–1823.
- Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4), 599–637.
- Wang, B., Cao, X., Theeuwes, J., Olivers, C. N. L., & Wang, Z. (2017). Separate capacities for storing different features in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(2), 226–236.
- Wang, L., Amalric, M., Fang, W., Jiang, X., Pallier, C., Figueira, S., ... Dehaene, S. (2019). Representation of spatial sequences using nested rules in human prefrontal cortex. *NeuroImage*, 186, 245–255.
- Webb, T. W., Sinha, I., & Cohen, J. D. (2021). Emergent symbols through binding in external memory. *ArXiv*. doi:10.48550/arXiv.2012.14601
- Weise, C., Ortiz, C. C., & Tibbetts, E. A. (2022). Paper wasps form abstract concept of “same and different.” *Proceedings of the Royal Society of London Series B: Biological Sciences*, 289(1979), 20221156. <https://doi.org/10.1098/rspb.2022.1156>
- Wood, J. N., & Wood, S. M. W. (2020). One-shot learning of view-invariant object representations in newborn chicks. *Cognition*, 199, 104192. doi:10.1016/j.cognition.2020.104192
- Xu, F. (2019). Toward a rational constructivist theory of cognitive development. *Psychological Review*, 126(6), 841–864.
- Xu, F., & Carey, S. (1996). Infants' metaphysics: The case of numerical identity. *Cognitive Psychology*, 30(2), 111–153.
- Xu, Y. (2017). Reevaluating the sensory account of visual working memory storage. *Trends in Cognitive Sciences*, 21(10), 794–815.
- Xu, Y. (2020). Revisit once more the sensory storage account of visual working memory. *Visual Cognition*, 5–8, 433–446.
- Xu, Y., & Vaziri-Pashkam, M. (2021a). Examining the coding strength of object identity and nonidentity features in human occipito-temporal cortex and convolutional neural networks. *Journal of Neuroscience*, 41(19), 4234–4252.
- Xu, Y., & Vaziri-Pashkam, M. (2021b). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, 12(2065), 1–16.
- Xu, Y., Zhou, X., Chen, S., & Li, F. (2019). Deep learning for multiple-object tracking: A survey. *IET Computer Vision*, 13(4), 355–368.
- Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3), 356–365.
- Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, 34(10), 515–525.
- Ye, X., & Durrett, G. (2022). The unreliability of explanations in few-shot in-context learning. *ArXiv*. doi:2205.03401
- Yildirim, I., & Jacobs, R. A. (2015). Learning multisensory representations for auditory-visual transfer of sequence category knowledge: A probabilistic language of thought approach. *Psychonomic Bulletin & Review*, 22(3), 673–686.
- Zettlemoyer, L. S., & Collins, M. (2005). Learning to map sentences to logical form: Structured classification with probabilistic categorical grammars. *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pp. 658–666.
- Zhou, K., Luo, H., Zhou, T., Zhuo, Y., & Chen, L. (2010). Topological change disturbs object continuity in attentive tracking. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50), 21920–21924.
- Zhu, Y., Gao, T., Fan, L., Huang, S., Edmonds, M., Liu, H., ... Zhu, S. C. (2020). Dark, beyond deep: A paradigm shift to cognitive AI with humanlike common sense. *Engineering*, 6(3), 310–345.

Open Peer Commentary

Never not the best: LoT and the explanation of person-level psychology

Louise Antony 

Department of Philosophy, University of Massachusetts, Amherst, MA, USA
lantony@umass.edu

doi:10.1017/S0140525X23001929, e262

Abstract

As Quilty-Dunn et al. observe, the language-of-thought hypothesis (LoTH) has fallen out of favor in philosophy. I will support the arguments made for its rehabilitation by Quilty-Dunn et al. by reviewing old, but still potent arguments for LoTH, and briefly criticizing recent proposed alternatives to LoT, such as Frances Egan's deflationism and Eric Schwitzgebel's dispositionalism, revealing inadequacies in such antirepresentational, anti-syntactic theories.

As noted by Quilty-Dunn et al., the language-of-thought hypothesis (LoTH) has fallen out of favor in philosophy. But why? I deeply appreciate the authors' careful and comprehensive review of contemporary empirical and mathematical work that supports the LoTH. I especially welcome their clarification of its core commitments, which enables us to see the LoTH at work in areas where its presence may not be apparent. But I feel that they are too concessive to LoT's critics. All of the considerations originally adduced in favor of the model still stand, particularly those that appeal to facts about person-level thought.

Understanding the propositional attitudes – believing, wanting, supposing, and so on – is one of the central goals of the philosophy-of-mind. What should constrain theorizing about them? Here are two surpassingly important *data*:

- (1) Propositional attitudes – particularly beliefs and desires – can combine to produce actions in ways that conform to what Aristotle called the “practical syllogism.”
- (2) A perfectly rational thinker can hold incompatible thoughts without realizing it.

As Jerry Fodor (1978) pointed out long ago, the hypothesis that propositional attitudes are functional states involving physically realized, syntactically structured representations offers smooth explanations of both these data. On the contrary, two leading anti-LoT theories cannot.

- (1) That the mental states of believing and desiring something can cause movements in the body is explained simply by their being realized physically – one doesn't need the LoT to do that. But not just any materialist theory can also explain why mental states can *rationalize* the movements they cause. If I want to snowshoe, and believe that I will be able to snowshoe if I go outside now, I will go outside now. That this belief-desire complex *makes rational* a certain course of

action is explained by the salutary formal relation among the structured representations that underlie the attitudes: If I want to X, and believe that if I do Y then I can accomplish X, then (*ceteris paribus*) I should do Y. The architecture of LoT, which guarantees the authors' properties 1, 2, 4, and 5, enables causal relations to track rational relations. (That the substituends of X and Y in [e.g.] the practical syllogism represent *propositions* is presupposed by the logic, although we'd need examples of different kinds of inference to secure property 3.)

- (2) Lois Lane believes that Superman can fly, and she believes that Clark Kent cannot. But Superman *is* Clark Kent, so Lois's beliefs conflict. But Lois is no dope; why can't she see the problem? According to the LoTH, it's because she has two lexically distinct representations for the same individual. Because her mental processes are sensitive only to the form of the (physically realized) representations, they have no compunction about placing *semantically* incompatible sentences into her belief box.

How can these data be explained without appeal to an LoT? I'll briefly discuss two views that are officially agnostic about the existence of language-like representations, but that hold that such representations need play no role in accounts of human mentality and behavior. They both fall short with respect to (1) and (2).

Dispositionalism, defended recently by Eric Schwitzgebel (2002), holds that believing (and desiring, etc.) is primarily a matter of being disposed to behave in certain characteristic patterns. A well-known problem for this view is specifying the pattern specific to a given belief. But there is a deeper problem. Beliefs do not *have* proprietary behavioral consequences – because they are *inferentially promiscuous*, they are willing to combine with *any* desire whatsoever to generate action. A dispositionalist can accommodate this point by saying that a signature behavioral profile is only determined holistically, by taking into account of all of an agent's beliefs and desires. But what does “taking into account” mean? The dispositionalist cannot rely on the logic of the practical syllogism to say what difference it would make to my behavior, given (say) my belief that it's snowing, whether I'm up for some wintry recreation or want to stay someplace warm.

Neither can the dispositionalist explain why the same set of motor movements is predictable when it's described one way, but not another. It's rational for me to, as I think of it, go outside. But going outside might, unbeknownst to me, involve stepping on a slippery surface. (I thought the walkway had been sanded.) My belief and desire rationalized my going outside, but they would not have rationalized my stepping on a slippery surface. The LoT explains why and how we do things “under a description,” as philosophers like to say (Antony, 1987).

Another view of belief, championed recently by Frances Egan (2014) is *deflationism*. Egan argues that the LoTH (a version of what she calls “Hyper Representationalism”) founders on the failure of efforts to give a naturalistic account of the representation relation. Characterizations of mental processes in terms of the manipulation of representations, she argues, should be viewed as merely useful “glosses,” not as serious posits of a mature cognitive science. Although I have to concede that we still don't know how to reduce intentionality to nonintentional conditions, I fault Egan for failing to recognize the role that the *syntactic* properties of LoT representations play in psychological dynamics. Whatever “Superman” *means*, its *lexical distinctness* from “Clark Kent” is sufficient to explain why Lois behaves differently when she deploys the first representation rather than the second.

To conclude: Why is the LoTH so unpopular? I suspect that it's because of a residual allegiance to behaviorism, with its commitment to empiricism (hence the enthusiasm for pattern-extraction models of thought, like deep neural networks [DNNs]), and its rejection of mentality as a genuine domain in nature. The person-level data stand on their own, but many thanks to the authors for demonstrating the utility, fecundity, and ubiquity of the LoTH in so many areas of contemporary psychology.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

- Antony, L. M. (1987). Attributions of intentional action. *Philosophical Studies*, 51(3), 311–323.
- Egan, F. (2014). How to think about mental content. *Philosophical Studies*, 170(1), 115–135.
- Fodor, J. A. (1978). Propositional attitudes. *The Monist*, 61, 501–523.
- Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs*, 36(2), 249–275.

Representational structures only make their mark over time: A case from memory

Sara Aronowitz 

Department of Philosophy, University of Toronto, Toronto, ON, Canada
s.aronowitz@utoronto.ca
<http://www-personal.umich.edu/~skaron/>

doi:10.1017/S0140525X23001905, e263

Abstract

Memory structures range across the dimensions that distinguish language-like thought. Recent work suggests agent- or situation-specific information is embedded in these structures. Understanding why this is, and pulling these structures apart, requires observing what happens under major changes. The evidence presented for the language-of-thought (LoT) does not look broadly enough across time to capture the function of representational structure.

The authors posit a single-format type, the language-of-thought (LoT), across both cognitive contexts and kinds of processing. Alongside asking whether this is true, we might also ask when it is true. In this commentary, I'll look at these two questions through the lens of long-term memory.

O'Keefe and Nadel (1978) made the case for a map-like format in the hippocampus, a structure distinct from the LoT. Although maps have symbolic elements, such as the graphical nature of a map of a subway route, they are also continuous and holistic in a way that LoT structures are not. I am unsure of whether the authors intended long-term memory to be a part of the swath of cognition covered by their theory, and to rehearse the map or language debate is not my aim here. I want instead to take

the hippocampal map hypothesis as a starting point to discuss two newer developments.

In their original book, O'Keefe and Nadel described a cognitive map that represents objective space abstracted from the creature's particular interactions with the environment. They note, for instance: "unlike the extra-hippocampal systems the locale system is relatively free from the effects of time and repetition" (p. 95). This idea of a map indifferent to the agent's path through it, and not greatly changed by repetition, shares some key similarities with the LoT format: Both can be updated quickly and categorically with new information, such as when I reorder a logical inference because of the introduction of a new proposition or remap a path based on the observation of a new obstacle.

In the case of memory, although the core idea of the hippocampus as a map is still popular, the notion of a fully objective map has been brought under strain. Work in reinforcement learning has proposed some of the computational work done by the hippocampus may employ a representation that is more path-dependent than a fully objective map: The successor representation (Dayan, 1993; Gershman, 2018; Momennejad et al., 2017). This representation stores expected (temporally discounted) connections between a state and the next states that will be visited. Using this representation in planning and learning is efficient. Standard successor learners can quickly figure out how to change course when they learn that rewards are redistributed (such as finding out a bet has doubled) but not when they learn that transitions between states have changed (such as finding out that pushing a button leads to a new floor). These learners make a compromise, easier computations at the cost of giving up some of the properties that LoT and map-like structures share.

Ormond and O'Keefe (2022) observe a different feature of hippocampal maps that seems to violate the idea that these maps are fully objective. During goal-directed behavior, they find that some place cells in rats represent their environment in a way that is distorted by increased firing when the animal is facing the goal, and incrementally decreased firing in directions farther from the goal. In a set of cells that were thought to represent invariant features of the environment, these patterns reflect goal location and shift during goal change.

In these two cases, researchers have theorized that initially indifferent and objective maps might encode some agent-centered information: Visited states and transitions in the successor representation, and goal location in the place cell subtypes. This leads to two connections with the LoT hypothesis.

First, although the authors acknowledge that the LoT is not the only format used in thought, they do not go far enough in thinking about how formats can be shifted and combined. In the case of map-like formats in memory, we've seen that the hypothesis that goal and path information seeps into the map does not mean a switch between discrete ways of representing but instead a sort of hybrid or intermediate representation. The successor representation is functionally in between a map-like model and a model-free algorithm, and Momennejad et al. go further to propose a hybrid successor/model-based learner. The goal-sensitive place cells are still place cells and are presumably used in navigation alongside the previously observed map-like structure.

If the six features noted by the authors are characteristic of one end of a spectrum of cognitive processing, noting that they obtain in various cases is not enough for thought to have a linguistic format. This is because such a system – not just one that comes in degrees of LoT-ishness, but one that instantiates these degrees in interlocking representations that work together – must be explained through the connections between formats, rather than

solely computations within a format. This renders the language-of-thought hypothesis somewhat toothless.

The second issue is about change in format over time. The authors take the relationship between more and less abstract representational systems to be one of realization, in the good case. Thinking about goal information in memory suggests an alternative: Even functional similarity may be a temporary and shallow equivalence.

If we start by looking at the cognition of a creature that has already learned a model of the world, and over a period where no substantial learning occurs, we should expect a period of equivalence between representational structures. But if this equivalence will not characterize how the representations were learned nor how they will shift and change, no more than we should expect a neural network that produces human-like behaviors to have acquired its expertise in a way even vaguely related to the way a human would. And once developed, the change, decay, or warping under pressure in these structures might again break the equivalence.

Memory structures, once hypothesized to be fully objective and agent-indifferent, now seem to contain some elements of goal or behavior dependence. The lesson of this for the LoT theory is that representational structures are ultimately distinguished in learning, decay, and other forms of change – not in stasis.

Competing interest. None.

References

- Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4), 613–624.
- Gershman, S. J. (2018). The successor representation: Its computational logic and neural substrates. *Journal of Neuroscience*, 38(33), 7193–7200.
- Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9), 680–692.
- O’Keefe, J. O., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Clarendon Press.
- Ormond, J., & O’Keefe, J. (2022). Hippocampal place cells have goal-oriented vector fields during navigation. *Nature*, 607(7920), 741–746.

Is evidence of language-like properties evidence of a language-of-thought architecture?

Nuhu Osman Attah^a and Edouard Machery^{a,b,c} 

^aDepartment of History and Philosophy of Science, University of Pittsburgh, Pittsburgh, PA, USA; ^bCenter for Philosophy of Science, University of Pittsburgh, Pittsburgh, PA, USA and ^cAfrican Centre for Epistemology and Philosophy of Science, University of Johannesburg, Johannesburg, South Africa
nuo2@pitt.edu, www.nuhuosmanattah.com
machery@pitt.edu, www.edouardmachery.com

doi:10.1017/S0140525X23001875, e264

Abstract

We argue that Quilty-Dunn et al.’s commitment to representational pluralism undermines their case for the language-of-thought hypothesis as the evidence they present is consistent with the operation of the other representational formats that they are willing to accept.

Quilty-Dunn et al. have convincingly shown that a variety of cognitive domains are characterized by some of the six properties they delineate: (1) Discrete constituents, (2) role-filler independence, (3) predicate–argument structure, (4) logical operators, (5) inferential promiscuity, and (6) abstract conceptual content. (We refer to these as the six “core properties.”) Foregrounding these properties is a worthwhile contribution because it establishes a framework and terminology for discussing features of cognition that hypotheses about representation must explain. We hope that this taxonomy will be expanded and refined: As it stands, some of the properties are defined so vaguely that they are too readily discoverable in cognition. For instance, role-filler independence requires that the same representational constituents can be deployed in different syntactic roles, but both the criterion for sameness of representational constituents and the relevant notion of syntax are left intuitive so that even the swapping of visual features of objects (e.g., misattributing the color of one object to another) counts as a demonstration of role-filler independence. On the other hand, the authors conveniently take successful demonstrations of compositionality in connectionist networks to fall short of role-filler independence because they “fail to preserve identity of the original representational elements” (target article, sect. 2, para. 7), even though no account of representational identity is given.

However, the authors’ aim is not merely to characterize these properties, but to show that they form the homeostatic cluster that marks a language of thought. This ambitious project fails because of the authors’ commitment to representational pluralism. The authors concede that language-like representations and the many other formats of representation that they are happy to accept share some properties: “Many, perhaps all, of these properties are not necessary for a representational scheme to count as a LoT, and some may be shared with other formats” (target article, sect. 2, para. 3). To give examples from the connectionist literature, even simple twentieth-century style connectionist networks form abstract representations (Clark, 1993) and modern networks fare even better (Stoianov & Zorzi, 2012); older networks can also bind values and variables (Smolensky, 1990); there has even been progress on the use of logical operators (Irving et al., 2016; Bansal, Loos, Rabe, Szegedy, & Wilcox, 2019; Dai, Xu, Yu, & Zhou, 2019). And in any case, because neural networks are universal function approximators there is ground for optimism about the prospects of architectures that do not implement a language of thought.

The authors’ representational pluralism undermines the inference to language-like representations from the observation of some of the core properties in some cognitive domain: These properties could simply result from any of the many other representational formats that the authors are willing to accept. The fallacy is similar to the issue with reverse inference (Machery, 2014; Poldrack, 2006): Although the likelihood of observing the core properties if representations are language-like is high, it is fallacious to infer that representations are language-like if these properties are observed because core properties could be observed even if representations are not language-like.

Quilty-Dunn et al. might reply that while some of the properties can be realized by nonlanguage-like representational formats, we are entitled to infer language-like representational structures where they cluster: As they say, such clustering “would be surprising from a theory-neutral point of view, but not from the perspective of LoTH” (target article, sect. 2, para. 13). We see two issues with this reply. First, only some

of the six core properties are observed in each of the few cognitive domains discussed in the paper: Three properties are demonstrated by implicit social cognition and four (the maximal number of cooccurring core properties) in the object-files case. Shall we conclude that only a few cognitive domains involve language-like representations? An interesting conclusion surely, but one that is much less exciting than the one touted by Quilty-Dunn et al.

Second, Quilty-Dunn et al. haven't even shown that clustering of the core properties is a unique prediction of the language-of-thought hypothesis. Many of the core properties are in fact coinstantiated in neural networks. For instance, the outputs of a sequence-to-sequence language model like BERT evince (at the very least) role-filler independence and predicate-argument structure (in addition to the general capacity for abstraction demonstrated by neural networks). Evidence suggests that these characteristics are underlain by systematic syntactic and semantic competences (Clark, Khandelwal, Levy, & Manning, 2019; Tenney, Das, & Pavlick, 2019). Thus, other architectures are consistent with the clustering of the core properties.

Perhaps the authors think that the burden-of-proof is on their opponents to show that these other formats exist and can account for the apparent clustering. But outside philosophy, such burden-of-proof claims are as weak an argument as it gets. Inferring a language-of-thought architecture on such shaky grounds also runs the risk of slowing research in computational neuroscience on new alternative cognitive architectures that are both neuroscientifically plausible and that can account for the core properties. Finally, and most important, alternatives to language-of-thought cognitive architectures have been investigated for decades, and the properties discussed by Quilty-Dunn et al. are known to result from these (Eliasmith, 2013; Eliasmith & Anderson, 2003; Smolensky, 1990, 1991). In none of these cases do the architectures merely implement a language-of-thought.

Finally, Quilty-Dunn et al. rely on the epistemic virtues of explanatory breadth and unification to support the language-of-thought hypothesis: As they say, "The chief aim [...] is to showcase LoTH's explanatory breadth and power in light of recent developments in cognitive science" (target article, sect. 1, para. 3). But an appeal to explanatory breadth runs against their pluralistic commitment: If the authors are serious about representational pluralism, it is hard to understand why they believe that explanatory breadth is a virtue or why any unification should be expected.

Although their defense of the language-of-thought hypothesis fails, Quilty-Dunn et al. are onto something important: We should expect cognition to exploit the core properties to solve some types of cognitive challenges, and we should thus predict their occurrence in some cognitive domains. Which tasks are facilitated by these properties and which life forms in the phylogenetic tree had to solve such tasks (and why) are exciting empirical questions.

Competing interest. None.

References

- Bansal, K., Loos, S., Rabe, M., Szegedy, C., & Wilcox, S. (2019). *HOList: An environment for machine learning of higher order logic theorem proving*. Proceedings of the 36th international conference on machine learning (Vol. 97, pp. 454–463).
- Clark, A. (1993). *Associative engines: Connectionism, concepts, and representational change*. MIT Press.
- Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). *What does BERT look at? An analysis of BERT's attention*. Proceedings of the 2019 ACL workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 276–286). Florence, Italy: Association for Computational Linguistics.

- Dai, W. Z., Xu, Q., Yu, Y., & Zhou, Z. H. (2019). *Bridging machine learning and logical reasoning by abductive learning*. Proceedings of the 33rd international conference on neural information processing systems (pp. 2811–2822). Red Hook, NY: Curran Associates Inc.
- Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.
- Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation and dynamics in neurobiological systems*. MIT Press.
- Irving, G., Szegedy, C., Alemi, A. A., Eén, N., Chollet, F., & Urban, J. (2016). DeepMath-deep sequence models for premise selection. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 2235–2243).
- Machery, E. (2014). In defense of reverse inference. *The British Journal for the Philosophy of Science*, 65, 251–267.
- Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, 10(2), 59–63.
- Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1–2), 159–216.
- Smolensky, P. (1991). Connectionism, constituency, and the language of thought. In B. M. Loewer & G. Rey (Eds.), *Meaning in mind: Fodor and his critics* (pp. 201–227). Oxford: Blackwell.
- Stoianov, I., & Zorzi, M. (2012). Emergence of a "visual number sense" in hierarchical generative models. *Nature Neuroscience*, 15(2), 194–196.
- Tenney, I., Das, D., & Pavlick, E. (2019). *BERT rediscovers the classical NLP pipeline*. Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 4593–4601). Florence, Italy: Association for Computational Linguistics.

Perception is iconic, perceptual working memory is discursive

Ned Block 

Department of Philosophy, New York University, New York, NY, USA
Ned.block@nyu.edu
<https://www.nedblock.us>

doi:10.1017/S0140525X23001899, e265

Abstract

The evidence that the target article cites for language-of-thought (LoT) structure in perceptual object representations concerns perceptual working memory, not perception. Perception is iconic, not structured like an LoT. Perceptual working memory representations contain the remnants of iconic perceptual representations, often recoded, in a discursive envelope.

In their wonderful and provocative target article, Quilty-Dunn et al. say perceptual object representations have language-of-thought (LoT) structure. However, there is plenty of evidence that perceptual object representations are iconic in a sense that excludes LoT representations; the evidence Quilty-Dunn et al. cite pertains to discursive *perceptual working memory* (WM) representations, not discursive *perceptual representations*. I will first present some evidence that perceptual object representations are iconic, then that WM representations are discursive. I will use the term "discursive" for representations that exhibit almost all of the six properties they cite rather than "LoT" because I doubt that even perceptual working memory exhibits all of them. (See Susan Carey's response to the target article.) But if there are no discursive representations in perception, there are no LoT representations either.

Apparent motion suggests iconic perceptual object representations. When two nearby objects flicker with the right parameters,

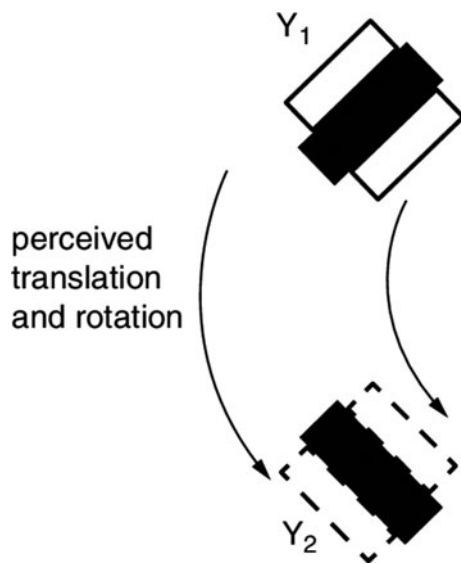


Figure 1 (Block). Items at Y1 and Y2 flicker so as to create apparent motion between them. The shapes are viewed in an apparatus in which a slightly different image is projected to each eye. This allows one version in which the black bar is part of a squarish shape and another version in which the bar protrudes, making the item look like an object instead of a shape. In the latter case, the subject sees rotation along with the movement but in the former case the subject sees just motion. Thanks to Ken Nakayama for providing this figure.

we see motion between them. Objects move while visible properties change gradually.

See the caption to [Figure 1](#). What suggests iconicity in this case of apparent motion is analog mirroring: Certain relations in the world are mirrored by representations that instantiate analogs of those relations in a way that is sensitive to degrees of difference. What is interesting about this case is that when the figure is perceived as an object, mirroring respects objecthood.

If the flickering objects are of different sizes, we see smooth expansion and contraction. One might suppose that the further apart the flickering objects are, the faster the rate of flicker would have to be to see motion. However, mirroring dictates the opposite because objects that are further apart take longer to traverse the distance. The further apart the flickering objects are, the longer the time span between flickers has to be to see motion (Korte's Third Law). The visual system prefers short motion paths between flickering objects but that preference is overridden if the shortest path involves biologically impossible motion (Shiffrar & Freyd, 1990) or if a moving object turns into an object of a different kind. In sum, perceptual representations are iconic in a sense that excludes discursive representations (see Block, 2023a, 2023b).

I now switch to the topic of perceptual working memory. Perceptual working memory often contains the remnants of perception – typically *not consciously experienced*. It can include iconic materials, but visual working memory often includes them in recoded form. One illustration of the partially nonperceptual nature of visual working memory is illustrated in [Figure 2](#). When the central disk and the donut surrounding it are presented simultaneously, there is center-surround suppression on the right, but not the left. However, when they are presented one at a time, with the first stimulus maintained in working memory, the collinear effect disappears (Bloem, Watanabe, Kibbe, & Ling, 2018). Thus a fundamental computational aspect of perception is absent in this working memory representation (see pp. 113–114 of Block, 2023a).

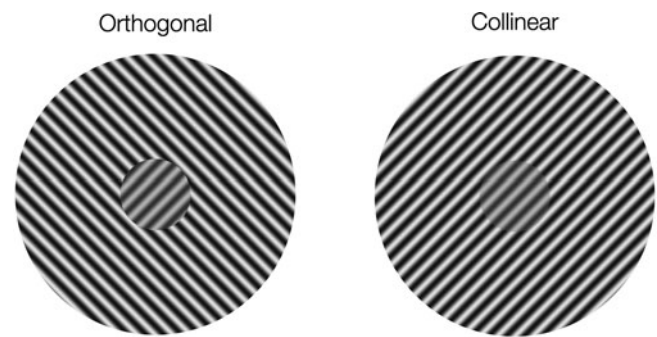


Figure 2 (Block). Central disk is the same on the left and on the right but it looks much higher in contrast on the left. The suppressive center/surround effect in the collinear case is because of a fundamental computation known as divisive normalization. Thanks to Sam Ling for providing this figure.

Quilty-Dunn et al.'s arguments for the iconic nature of perception involve the “object files” of perceptual working memory. But working memory representations in visual areas are often recoded outside of the classic visual system, for example, in the intraparietal sulcus, while they disappear from visual cortex because of ongoing visual stimulation (Rademaker, Chunharas, & Serences, 2019). Perceptual information is often recoded in the service of specific tasks. Kwak and Curtis (2022) showed subject clouds of moving dots and also oriented gratings, asking them to remember the directions. They found that brain decoding on either of these working memory representation worked on the other suggesting that working memory coded what was in common to the two kinds of percepts, eliminating the moving dots and the gratings, replacing them with representations of vectors, showing that many iconic features can be altered or discarded in working memory.

Of course perceptual working memory is constantly interacting with perception. Quilty-Dunn (2023) argues convincingly that this interaction is crucial to longer term perceptions, for example, perceptions that span saccades. As Quilty-Dunn notes, perception does not start anew after each saccade, so there must be some perceptual – I would say iconic – information preserved by the saccade. True, but there is plenty of evidence for at least some loss of iconicity in transsaccadic memory (summarized in Block, 2023a, pp. 261–262). For example, in the famous Sperling effect, a multirow array of letters is presented briefly, but a cue presented after the stimulus stops can focus attention on any one of the rows, allowing reporting of all or almost all the items. However, if the array is presented before the saccade and the cue presented afterward, the Sperling effect disappears, showing that transsaccadic memory can erase the iconic memory that the Sperling effect depends on.

Quilty-Dunn describes a perceptual effect known as the motion repulsion illusion. If dots moving in one direction are superimposed on dots moving in a different direction, the perceived angle between the two directions is exaggerated in perception. Kang et al. showed that the same effect occurs if one set of moving dots is seen while the other is held in working memory (Kang, Hong, Blake, & Woodman, 2011). This result suggests that there are iconic elements in perceptual working memory but does nothing to show that perception is not iconic.

As Quilty-Dunn notes, perception can distort working memory and conversely. Teng and Kravitz (2019) showed that colors and orientations in each of perception and working memory affect the other, commenting that this is no doubt

because of overlapping representations in visual processing areas. However, as the authors note, these results are compatible with the involvement of prefrontal cortex in working memory. The overlap of sensory coding between visual working memory and vision does not preclude partial reformatting in working memory or the inclusion of iconic information in a discursive envelope.

In sum, perceptual working memory can preserve some aspects of iconic perceptual representation even if it includes it in a discursive envelope. Quilty-Dunn et al.'s results depend on the discursive envelope, not the iconic perceptual representations.


Financial support. The author acknowledges Templeton World Charities for support.

Competing interest. None.

References

- Block, N. (2023a). *The border between seeing and thinking*. Oxford University Press. Open access at <https://global.oup.com/academic/product/the-border-between-seeing-and-thinking-9780197622223?cc=us&lang=en&>
- Block, N. (2023b). Let's get rid of the concept of an object file. In B. McLaughlin & J. Cohen (Eds.), *Contemporary debates in philosophy of mind* (pp. 494–516). Wiley Blackwell.
- Bloem, I. M., Watanabe, Y. L., Kibbe, M. M., & Ling, S. (2018). Visual memories bypass normalization. *Psychological Science*, 29(5), 845–856. doi:10.1177/0956797617747091
- Kang, M.-S., Hong, S. W., Blake, R., & Woodman, G. F. (2011). Visual working memory contaminates perception. *Psychonomic Bulletin & Review*, 18(5), 860–869. doi:10.3758/s13423-011-0126-5
- Kwak, Y., & Curtis, C. E. (2022). Unveiling the abstract format of mnemonic representations. *Neuron*, 110(11), 1822–1828. doi:<https://doi.org/10.1016/j.neuron.2022.03.016>
- Quilty-Dunn, J. (2023). *Remnants of perception: Comments on Block*. Paper presented at the American Philosophical Association, San Francisco, April 6, 2023.
- Rademaker, R. L., Chunharas, C., & Serences, J. T. (2019). Coexisting representations of sensory and mnemonic information in human visual cortex. *Nature Neuroscience*, 22(8), 1336–1344. doi:10.1038/s41593-019-0428-x
- Shiffrar, M., & Freyd, J. J. (1990). Apparent motion of the human body. *Psychological Science*, 1(4), 257–264. doi:10.1111/j.1467-9280.1990.tb00210.x
- Teng, C., & Kravitz, D. J. (2019). Visual working memory directly alters perception. *Nature Human Behaviour*, 3(8), 827–836. doi:10.1038/s41562-019-0640-4

Natural logic and baby LoTH

Irene Canudas-Grabolosa^a, Ana Martín-Salguero^{b,c}
and Luca L. Bonatti^{b,d} 

^aDepartment of Psychology, Department of Linguistics, Harvard University, Cambridge, MA, USA; ^bCenter for Brain and Cognition, Universitat Pompeu Fabra, Barcelona, Spain; ^cCognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin Center, Gif/Yvette, France and ^dICREA, Barcelona, Spain

irenecanudas@gmail.com, ana.martin@upf.edu, lucabonatti@mac.com

doi:10.1017/S0140525X23001942, e266

Abstract

Language-of-thought hypothesis (LoTH) is having a profound impact on cognition studies. However, much remains unknown about its basic primitives and generative operations. Infant studies are fundamental, but methodologically very challenging. By distilling potential primitives from work in natural-language semantics, an approach beyond the corset of standard formal logic may be undertaken. Still, the road ahead is challenging and long.

Fodor had the gift of conceiving extremely simple ideas with extremely deep and rich consequences. Language-of-thought hypothesis (LoTH) is perhaps the best, but not the unique, example of this gift. Quilty-Dunn et al.'s article is a very forceful testimony of how lively and far-reaching LoTH is. The very fact that they use a cluster of properties that prescind from most traditional arguments for LoT is in itself a proof of its richness. At the same time, as it is clear in the target article, like other cases (modularity witness it), Fodor's LoTH was more a research program than an hypothesis; in his words, it's probably a genus, but, we would add, one whose actual species are still barely known. This dearth of knowledge is particularly acute for one of the fundamental issues in characterizing LoT(s): Identify the basic primitives available endogenously in human thinking. In adults, a recent work investigated modular LoTs defined over various domains, proposing primitives and compositional routines (Al Roumi, Marti, Wang, Amalric, & Dehaene, 2021; Dehaene, Al Roumi, Lakretz, Planton, & Sablé-Meyer, 2022; Planton et al., 2021; Sablé-Meyer et al., 2021; Sablé-Meyer, Ellis, Tenenbaum, & Dehaene, 2022). However exciting and important to characterize human singularity, these theories do not clarify the origins of LoT or its role in general human cognition. They can check out all the list of properties in Quilty-Dunn et al.'s cluster, and yet remain confined to the specific domain they have been tested, in adults. They are compatible with the fact that language interactions, or instruction, contributes to their appearance.

Although there is little doubt that when linguistic competence kicks in, human language competence is explained by reference to a system of structures encompassing many properties of a general LoT, the crucial open questions are whether properties of general thinking are somehow imported from linguistic structures, as many would hold (Carruthers, 2002; Spelke, 2003), or else are inherent properties of the mind, and if they encompass the logical concepts that make LoT cross-domain and compositional. Progress on these questions can be achieved by investigating the existence and nature of the logical primitives available to preverbal infants, who are likely not affected by instructions or massive language experience. Unfortunately, these investigations can be counted on the fingers of one hand. They are also very difficult, as they require creating scenes deprived of verbal cues that likely embed logical inferences, something that at best can be supported by arguments to the best explanation. We thank Quilty-Dunn et al., who agree with us that a baby-LoTH has the upper hand relative to alternative theories, but they are more optimistic than us: Alternative explanations, perhaps compatible, perhaps incompatible with some declination of LoTH (Leahy & Carey, 2020), exist and have to be addressed experimentally. Furthermore, an unified explanation of the early putative indications of logical thinking (Cesana-Arlotti et al., 2018; Cesana-Arlotti, Kovács, & Téglás, 2020; Cesana-Arlotti, Téglás, & Bonatti, 2012; Cesana-Arlotti, Varga, & Téglás, 2022) and the later failures at making action plans consistent with it (Feiman, Mody, & Carey, 2022; Leahy, Huemer, Steele, Alderete, & Carey, 2022; Mody & Carey, 2016) is still missing. All these issues require painstaking research.

As baby LoTH supporters, we believe that the most serious question remains the identification of a plausible repertoire of early LoT primitives. Short of the success in disjunctive reasoning (Cesana-Arlotti et al., 2018), little exists about other logical components of an LoT, while some arguably plausible candidates – for

example, simple relations such as “Same/Different” – do not seem to be supported (Hochmann, 2022; Hochmann et al., 2017; Hochmann, Mody, & Carey, 2016). Where to look for plausible candidates? The naive approach we, our collaborators, and others have taken has been to follow the guidance of formal logic, hence looking for connectives, quantifiers, or Boolean concepts. This is a plausible approach, but deep down is based on the arbitrary assumption that the forms of human thoughts comply with descriptions largely developed for mathematical elegance rather than for psychological reality. Whether thought fits the well-defined but rigid corset of logical systems is far from granted. Indeed, even with the baby LoT case discussed by Quilty-Dunn et al., which we contributed to develop (Cesana-Arlotti et al., 2018), it is still not clear whether infants’ mental representations involve disjunctions (A or B), possible alternatives (maybe A, maybe B), or quantified representations (“unknown x”).

Another possible approach is to invert the relation between natural languages and thought. Rather than regarding them as the origin of logical abilities in thought, one could look at their semantics as crystalized repositories of thought primitives. Under this perspective, LoT primitives may well differ from those familiar from logic. For example, all studied languages contain polarity items, but these have no place in logical systems. Yet, undoubtedly, their behavior is “logical”; perhaps they signal the presence of primitive logical operations that are at the source of their widespread presence in natural languages. Likewise, operations such as exhaustification, which seems to be necessary to explain many patterns of implications in language (Chierchia, 2013), do not exist in logic, but could potentially be present in an LoT, as a primitive operation defined over sets and set relations (another area which, with few exceptions [Feigenson & Halberda, 2004, 2008; Zosh, Halberda, & Feigenson, 2011], is still poorly known). Tense and modal structure can also offer case studies for the identification of potential logical primitives.

We feel that the interaction between natural-language semantics and psychology can be a fruitful way to unlock basic potential primitives of thought. From there on, painful and long case studies have to be developed to trace back their nonverbal origins in infant LoT(s). It took about 50 years to transform LoTH into a fruitful research program; we bet that it won’t take much less to go from the acceptance of the genus LoT to the discovery of its species and their common origins in preverbal infants. Nonetheless, we feel, the potential for new insights and discoveries makes this endeavor worthwhile undertaking.

Financial support. This work was supported by the Ministerio de Ciencia e Innovación Grant PID2019-108494GB-I00 to L. L. B.

Competing interest. None.

References

- Al Roumi, F., Marti, S., Wang, L., Amalric, M., & Dehaene, S. (2021). Mental compression of spatial sequences in human working memory using numerical and geometrical primitives. *Neuron*, 109(16), 2627–2639, e4.
- Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, 25(6), 657–726.
- Cesana-Arlotti, N., Kovács, Á. M., & Téglás, E. (2020). Infants recruit logic to learn about the social world. *Nature Communications*, 11(1), 1–9.
- Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science (New York, N.Y.)*, 359(6381), 1263–1266.

- Cesana-Arlotti, N., Téglás, E., & Bonatti, L. L. (2012). The probable and the possible at 12 months: Intuitive reasoning about the uncertain future. *Advances in Child Development and Behavior*, 43, 1–25.
- Cesana-Arlotti, N., Varga, B., & Téglás, E. (2022). The pupillometry of the possible: An investigation of infants’ representation of alternative possibilities. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 377(1866), 20210343.
- Chierchia, G. (2013). *Logic in grammar: Polarity, free choice, and intervention*. MIT Press.
- Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: A hypothesis about human singularity. *Trends in Cognitive Science*, 26(9), 751–766.
- Feigenson, L., & Halberda, J. (2004). Infants chunk object arrays into sets of individuals. *Cognition*, 91(2), 173–190.
- Feigenson, L., & Halberda, J. (2008). Conceptual knowledge increases infants’ memory capacity. *Proceedings of the National Academy of Sciences of the United States of America*, 105(29), 9926–9930.
- Feiman, R., Mody, S., & Carey, S. (2022). The development of reasoning by exclusion in infancy. *Cognitive Psychology*, 135, 101473.
- Hochmann, J. R. (2022). Representations of abstract relations in infancy. *Open Mind*, 6, 291–310.
- Hochmann, J. R., Mody, S., & Carey, S. (2016). Infants’ representations of same and different in match- and non-match-to-sample. *Cognitive Psychology*, 86, 87–111.
- Hochmann, J.-R., Tuerk, A. S., Sanborn, S., Zhu, R., Long, R., Dempster, M., & Carey, S. (2017). Children’s representation of abstract relations in relational/array match-to-sample tasks. *Cognitive Psychology*, 99, 17–43.
- Leahy, B., Huemer, M., Steele, M., Alderete, S., & Carey, S. (2022). Minimal representations of possibility at age 3. *Proceedings of the National Academy of Sciences of the United States of America*, 119(52), e2207499119.
- Leahy, B. P., & Carey, S. E. (2020). The acquisition of modal concepts. *Trends in Cognitive Sciences*, 24(1), 65–78.
- Mody, S., & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, 154, 40–48.
- Planton, S., van Kerkoerle, T., Abbih, L., Maheu, M., Meyniel, F., Sigman, M., ... Dehaene, S. (2021). A theory of memory for binary sequences: Evidence for a mental compression algorithm in humans. *PLoS Computational Biology*, 17(1), e1008598.
- Sablé-Meyer, M., Ellis, K., Tenenbaum, J., & Dehaene, S. (2022). A language of thought for the mental representation of geometric shapes. *Cognitive Psychology*, 139, 101527.
- Sablé-Meyer, M., Fagot, J., Caparos, S., van Kerkoerle, T., Amalric, M., & Dehaene, S. (2021). Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity. *Proceedings of the National Academy of Sciences of the United States of America*, 118(16), e2023123118.
- Spelke, E. S. (2003). What makes us smart? Core knowledge and natural language. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 277–311). MIT Press.
- Zosh, J. M., Halberda, J., & Feigenson, L. (2011). Memory for multiple visual ensembles in infancy. *Journal of Experimental Psychology: General*, 140(2), 141–158.

Do nonlinguistic creatures deploy mental symbols for logical connectives in reasoning?

Susan Carey 

Department of Psychology, Harvard University, Cambridge, MA, USA
scarey@wjh.harvard.edu
<https://www.harvardlds.org/our-labs/carey-lab/susan-carey/>

doi:10.1017/S0140525X23001917, e267

Abstract

Some nonlinguistic systems of representation display *some* of the six features of a language-of-thought (LoT) delineated by Quilty-Dunn et al. But they conjecture something stronger: That all six features cooccur homeostatically in nonlinguistic thought. Here I argue that there is no good evidence for nonlinguistic deductive reasoning involving the disjunctive syllogism. Animals and pre-linguistic children probably do not make logical inferences.

In the landmark target article, Quilty-Dunn et al. establish that object perception and visual working memory both display *some* of their six features of a language-of-thought (LoT). Nonetheless, I believe evidence to date fails to make a convincing case for predicate–argument structure and logical connectives in nonlinguistic thought. Here, I examine Quilty-Dunn et al.’s evidence for nonverbal disjunctive syllogism inferences.

Who counts as prelinguistic? Children master the basic argument structure of English sentences between 12 and 15 months of age (Fisher, Jin, & Scott, 2019). For French and Hungarian, children have mastered the logical meanings of words for “not” and “no” at least by 17 or 18 months of age (e.g., de Carvalho, Crimon, Barrault, Trueswell, & Cristophe, 2021). So only children under 17 months of age can be safely considered prelinguistic with respect to propositional representations and negation.

What counts as a *mental symbol* for disjunction or negation? The numerical content in visual working memory models of small sets of explicitly represented individuals, which support 1–1 correspondence operations is *implicit*. Concepts can also be “proto” or “precursor” versions of later emerging ones, expressing *part* of some target logical function, but not the whole function (as analog number representations are proto-integer concepts).

All animal thought is nonlinguistic. Regarding the disjunctive syllogism, Quilty-Dunn et al. appeal first to Call’s two-cup task, which is solved by *some* individuals of many species. However, except for adult great apes, often half or more of individuals tested fail even after hundreds of trials of training (e.g.,

Ferrigno, Huang, & Cantlon, 2021). Prelinguistic infants (i.e., 15-month olds) all dramatically fail the two-cup task. When asked to “find the toy,” they choose the empty container 50% of the time, failing to eliminate an option upon learning it is empty (a *contrary* of containing the reward) or upon learning that it does *not* contain the hidden reward. Importantly, control experiments showed that the 15-month olds were not confused by the actions that revealed one container empty, had not forgotten the reward, and wanted it. Further, the literature on 10- to 12-month olds’ working memory argues against other plausible hypotheses about task demands that might be masking competence. In contrast, 17-month olds succeeded, needing no

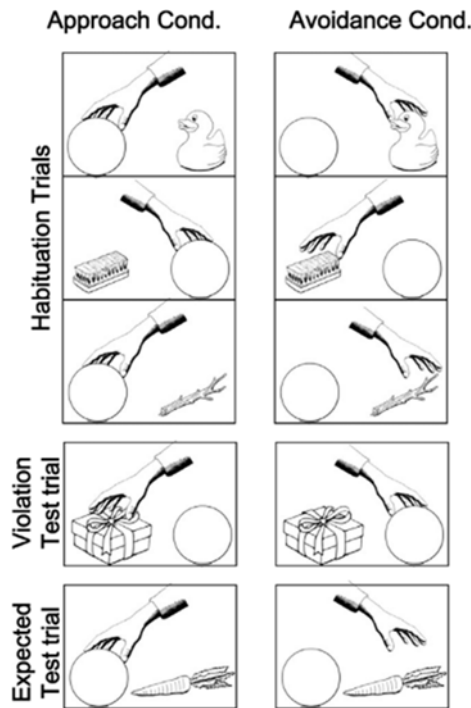


Figure 1 (Carey). During habituation in the approach condition, an agent repeatedly reaches for A (a ball in this example) over B, C, D,.... In the avoidance condition the agent repeatedly reaches for whatever is not A (here, a duck, a brush, a stick, a present, a carrot, etc.). The test trials establish that the child extends this pattern to a new pair A X, for example, a ball and a car, generalizing habituation when the agent’s action matches the pattern seen up to then (i.e., reaching to the ball in the approach condition and reaching to the car in the avoidance condition) and recovering interest if the agent’s action violates the pattern (from Feiman, Carey, & Cushman, 2015).

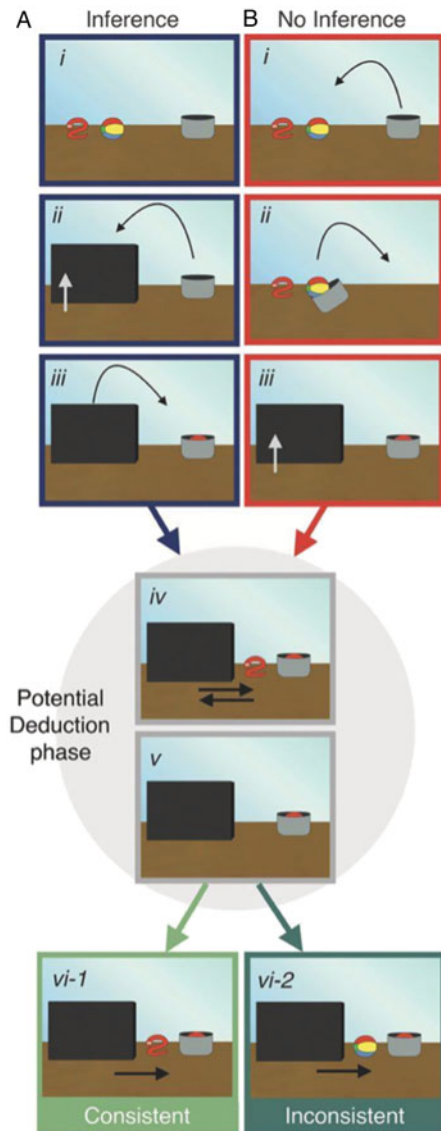


Figure 2 (Carey). Cesana-Arlotti et al.’s (2018) object disambiguation paradigm. **Focus on inference condition (A):** During the setup, the child sees a ball and a snake on the stage, these are occluded, and a cup swoops in and picks up one of them (unknown which one). In the potential deduction phase, one of the objects comes out from behind the screen and returns (here the snake). During the test phase, the child recovers interest either if the ball emerges from behind the occluder (shown here, inconsistent event) or if the snake is revealed in the cup (not shown here, but the inconsistent event for another group of infants). (B): in the No Inference condition, not discussed here, the child knows where the snake and the ball are from the beginning of the experiment.

training, at which age they show a domain-general capacity to eliminate options in indirect screening off trials in causal reasoning as well (Feiman, Mody, & Carey, 2022). But 17-month olds are not prelinguistic creatures.

There is convergent evidence for the absence of negation in prelinguistic human thought (Fig. 1). Fourteen-month-old infants learn that the agent likes or wants A in the approach condition, but fail to learn the agent does *not* like A, or wants anything that is *not* A in the avoidance condition.

Quilty-Dunn et al. draw on Cesana-Arlotti et al. (2018; Fig. 2) as evidence for reasoning involving the disjunctive syllogism by prelinguistic (12-month olds) infants. The flip side of worrying about task demands explaining failures is worrying about spurious successes. In this case, the well-characterized object file visual working memory system fully explains all the data. When small numbers of attended objects are occluded, working memory models with one object file for each attended object, with property and kind information bound to each, are formed and maintained as the event unfolds. Those representations support reidentifying an object when it comes back into view.

Infants make mental models of objects that are occluded (a snake and a ball; Fig. 2). This model is held in working memory as the child watches the unfolding scene. The child only updates that working model with respect to further information when that information becomes available, and only reacts with surprise if

they see something inconsistent with that model. When the cup swoops up one of the objects, this is still consistent with the snake and the ball being occluded in the scene. When the child sees the snake comes out from behind the occluder, the location of the snake is added to the model; a 1–1 mapping computation between the objects in the scene and those in working memory completes the model. Importantly in this process the child need not wonder whether the object in the object in the cup is the ball *or* the snake, nor ever draws a conclusion from the fact that the snake is *not* the ball. There need be no implicit or proto concepts of negation or disjunction in this process. This 1–1 mapping computation predicts the same pupil dilation and looking patterns during the potential deduction phase as does the disjunctive syllogism hypothesis and could underlie all of the successes attested in children under 17 months of age to date.

With respect to nonhuman animals, Quilty-Dunn et al. mainly draw on above chance performance in the three- and four-cup tasks (Fig. 3), first introduced by Mody and Carey (2016). The developmental facts are now clear from many replications (Leahy, Huemer, Steele, Alderette, & Carey, 2022). At 2.5 years of age children choose the certain cup exactly half of the time, on both the three- and four-cup tasks, and by age 3, 80% still choose it half of the time, whereas the remaining children always choose the certain cup. Even 50% is better than chance (33% on both tasks, because even 17-month olds can eliminate an option).

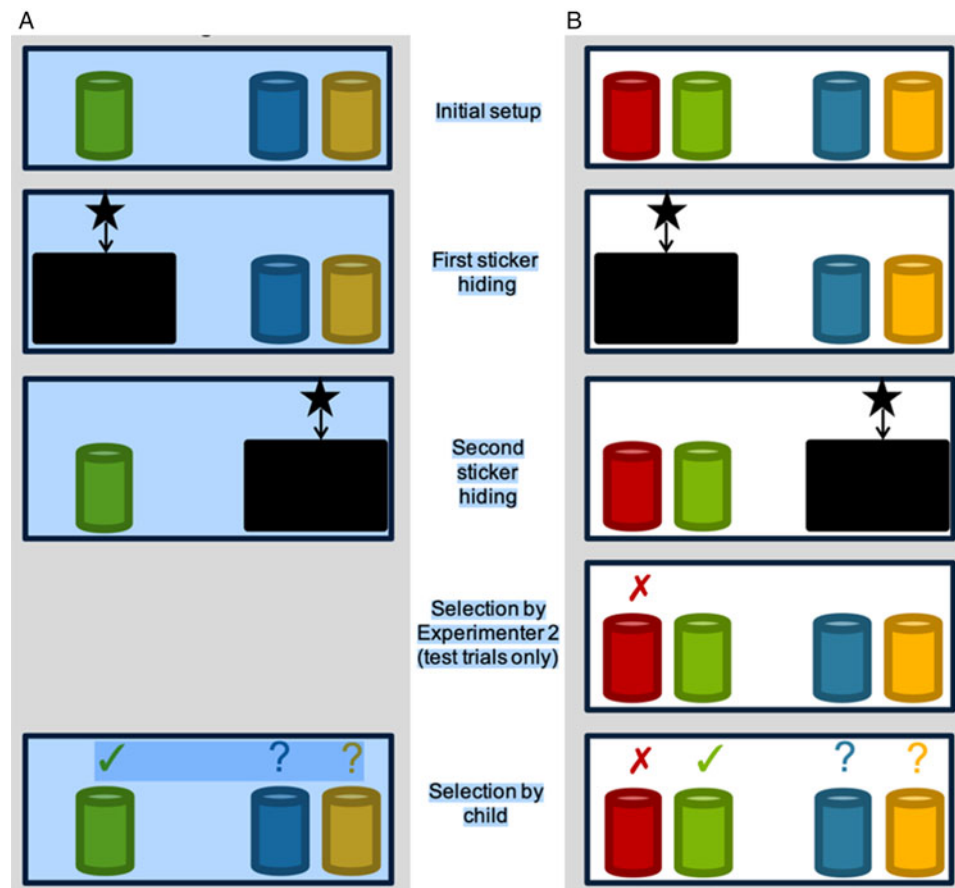


Figure 3 (Carey). (A) Three-cup procedure; child has one choice to obtain a sticker. The green cup is certain to have a prize; a prize can be in either the blue or yellow cup. (B) Four-cup procedure; the child again has one choice to obtain a sticker, and again a prize can be in either the blue or yellow cup, whereas the green cup is certain to contain a prize if the child can eliminate the possibility of red's containing a sticker upon seeing the red cup is empty. The pattern of response that maximizes reward is to always choose the green cup in both three- and four-cup trials.

Leahy et al. (2022) establishes that this 50% performance is because of a proto-concept of possibility – children simulate one possible location on the doubleton side and take that simulation as equivalent to the simulated location on the singleton side. They do not draw on a representation OR, or on a representation POSSIBLE, and thus are not reasoning through the disjunctive syllogism. Engelmann et al. (2022), in a later paper than that discussed by Quilty-Dunn et al. (Engelmann et al., 2021), find that chimpanzees pattern exactly with 2.5-year olds, choosing the singleton cup exactly half of the time in both the three- and four-cup tasks, and conclude they are probably not reasoning deductively.

Although I agree with Quilty-Dunn et al. that possibility of logical connectives in nonlinguistic reasoning is still open, the evidence for the disjunctive syllogism that they cite does not show logical connectives in nonlinguistic thought.

Financial support. Funding for research reported here from my lab was provided by Harvard University – support for graduate students and their research, internal grants for faculty research – and by a Network Grant from the McDonnell Foundation (“The Ontogenetic Origins of Abstract Combinatorial Thought”). I am extremely grateful to both institutions for their support of this work.

Competing interest. None.

References

- Cesana-Arlotti, N., Martin, A., Teglas, A., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science (New York, N.Y.)*, 359, 1263–1266.
- de Carvalho, A., Crimon, C., Barrault, A., Trueswell, J., & Christophe, A. (2021). “Look. It is not a bamoule”: 18- and 24-month-olds can use negative sentences to constrain their interpretation of novel word meanings. *Developmental Science*, 24(4), e13085.
- Engelmann, J., Haux, L., Völter, C. J., Call, J., Rakoczy, H., Herrmann, E., & Schleihauf, H. (2022). Do chimpanzees reason logically? *Child Development*, 94, 1–15. <https://doi.org/10.1111/cdev.13861>
- Engelmann, J., Volter, C. J., O'Madagain, C., Prot, M., Haun, D. B., Rakoczy, H., & Herrmann, E. (2021). Chimpanzees consider alternative possibilities. *Current Biology*, 32, R1–R3.
- Feiman, R., Carey, S., & Cushman, F. (2015). Infants' representations of others' goals: Representing approach over avoidance. *Cognition*, 136, 204–214.
- Feiman, R., Mody, S., & Carey, S. (2022). The development of reasoning by exclusion in infancy. *Cognitive Psychology*, 135, 101473, 1–20.
- Ferrigno, S., Huang, Y., & Cantlon, J. F. (2021). Reasoning through the disjunctive syllogism in monkeys. *Psychological Science*, 32, 292–300.
- Fisher, C., Jin, K., & Scott, R. M. (2019). The developmental origins of syntactic bootstrapping. *Topics in Cognitive Science*, 12, 48–77.
- Leahy, B., Huemer, M., Steele, M., Alderette, S., & Carey, S. (2022). Minimal representations of possibility at age 3. *Proceedings of the National Academy of Sciences of the United States of America*, 119, e2207499119.
- Mody, S., & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, 154, 40–48.

The reemergence of the language-of-thought hypothesis: Consequences for the development of the logic of thought

Nicolò Cesana-Arlotti 

Department of Psychology, Yale University, New Haven, CT, USA
nicolo.cesana-arlotti@yale.edu; www.nicolocesanaarlotti.com

doi:10.1017/S0140525X23001802, e268

Abstract

Quilty-Dunn et al. defended the reemergence of language-of-thought hypothesis (LoTH). My commentary builds up implications for the study of the development of our logical capacities. Empirical support for logically augmented LoT systems calls for the investigation of their logical primitives and developmental origin. Furthermore, Quilty-Dunn et al.'s characterization of LoT helps the quest for the foundation of logic by dissociating logical cognition from natural language.

The connections between language-of-thought (LoT), learning, and the development of logic were central in Fodor's proposal (Fodor, 1979). He pointed out that efficient learning by hypothesis-confirmation requires combinatorial, structured representations. Quilty-Dunn et al.'s article vindicates Fodor's conjecture: Contemporary cognitive science confirms that human-like flexibility and systematicity in learning (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Goodman, Tenenbaum, & Gerstenberg 2015; Piantadosi, Tenenbaum, & Goodman, 2016), and the ability to master a natural language (Chierchia, 2013; Pietroski, 2018), are best explained by LoT-like cognitive systems *augmented* with a repertoire of logical operators.

Fodor also argued that the *compositional* logical primitives of LoT (the logical building blocks that are not decomposed in more basic operators) must be *developmental* primitives – representations that are not learned – because concept learning requires decomposition. To be sure, we can “decompose” logical notions. But to do so, we need an equivalent or more powerful (expressive) logic. For instance, the operators of propositional logic can be interdefined (e.g., “p OR q” = “IF NOT p THEN q”) or can be defined by more expressive logical systems (e.g., lambda calculus or combinatory logic; Piantadosi, 2021). So, although children and adults could learn specific logical notions, this would require a LoT with equivalent, or more powerful, logical primitives.

As a result, the reemergence of LoTH *carries important consequences* for the study of the development of logic in the mind. If human cognition traffics in logically rich LoT systems, then cognitive development must start with a firm foundation of primitive logical capacities. But if not learning, what is the origin of our logical primitives? And what natural logical resources are in place when learning begins?

My next point expands on the hypothesis that natural language may not be the unique source of our logical capacities. I fully agree with Quilty-Dunn et al. that serious consideration should be given to the alternative picture: Logical primitives might be in place in LoTs distinct from the natural language. First, preverbal logical representations can explain our capacity to acquire many logical concepts through language and acculturation. Second, preverbal logic might play a role in accounting for infants' surprising learning potentials (Cesana-Arlotti, Kovács, & Téglás, 2020). After all, logically augmented LoTs are powerful hypothesis-testing devices.

With my collaborators, we have begun to investigate preverbal infants' logical abilities, targeting disjunction, the logical relation between two or more representations entailing that at least one of them is true (expressed by “OR” in formal logic). We tested whether 12-month-olds represented the identity of a half-hidden object compatible with two possibilities and inferred one identity based on evidence incompatible with the other (Figure 1A).

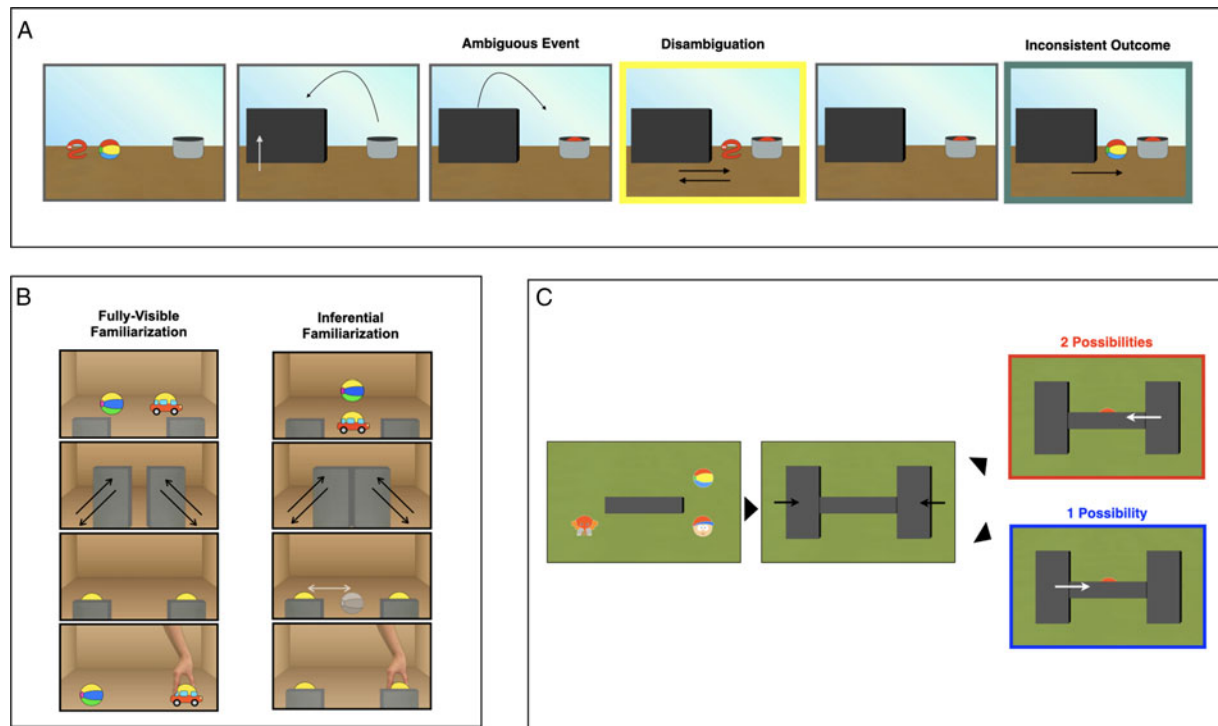


Figure 1 (Cesana-Arlotti). Tests of infants' disjunctive representation. (A) Infants are presented with movies where a half-hidden object is compatible with just two possible identities (ambiguous event; the half-hidden object is the snake OR the ball). Next, infants see which object is outside the cup, evidence that rules out one of the alternatives (disambiguation; the snake is *outside* the cup; so, the hidden object is the ball). We found that infants reacted to the disambiguating evidence with a reorientation of attention toward the half-hidden object, and then were surprised (i.e., look longer) by a later violation of the logical expectation that the other object is inside the cup (inconsistent outcome; the ball is outside the cup). Importantly, higher attentional reorientation at the time of the exclusion was predictive of later surprise. At the same time, this relationship was absent in a noninferential control condition (adapted from Cesana-Arlotti et al., 2018). (B) Infants were familiarized with a choice of an object which was either directly visible (fully visible familiarization) or had to be inferred via disjunctive inference (inferential familiarization). Infants familiarized via inference performed just as well as those who could directly see the choice (adapted from Cesana-Arlotti et al., 2020). (C) Infants watched visually identical events where a half-hidden object was compatible with a varying number of identities (one or two). Infants' pupil dilation (an index of processing load) was higher when there were two alternatives compatible with the object, suggesting that infants were not simulating just a single identity regardless of the alternative possibilities (adapted from Cesana-Arlotti et al., 2022).

Infants' looking times and oculomotor responses provided evidence of disjunctive representation at the preverbal stage (Cesana-Arlotti et al., 2018).

Evidence of preverbal logical abilities calls for a challenge: Extending the notion of logical representation beyond language and its formalizations (Bermúdez, 2007; Burge, 2010). To formulate and test hypotheses about the presence and nature of preverbal logical representations, we need a framework that could dissociate logic from language. Quilty-Dunn et al.'s homeostatic characterization of LoT offers a tool to this end: Unlike natural language, preverbal infants' logical primitives might not have all the properties of the LoT format. To conclude my commentary, I ask whether we have reasons to think infants' disjunctive inferences have key properties of an LoT.

First, infants' disjunctive inference displays evidence of *inferential promiscuity*. We found that infants quickly learn the preference of an agent reaching for a hidden goal, which has to be identified by exclusion (Cesana-Arlotti et al., 2020). The infants who had to learn the preference based on the inference (experiment 4) performed at the same level as those who could directly see the agent's goal (experiment 3, Figure 1B). This is striking given the few demonstrations needed by infants and the previous finding that observing few inconsistent reaches is sufficient to disrupt their learning (Luo, Hennefield, Mou, vanMarle, & Markson, 2017). New experiments should systematically investigate whether the disjunctive representations deployed by infants and adults in

processing visual scenes trigger automatic inferences (Braine & O'Brien, 1998; Quilty-Dunn & Mandelbaum, 2018).

Second, it is currently unclear whether infants' disjunctive representations have *discrete constituency*. Operators of formal logic are discrete symbols (e.g., "p OR q," "NOT p") that encode binary or monadic logical relations (e.g., truth-functions). Discrete constituency is crucial for formal logic because it supports compositionality: New logical relations can be expressed by embedding logical operators (e.g., "NOT (p OR q)"). Unlike formal logic, preverbal disjunctive inferences might use a format with no discrete logical operators, like the mental model theory (Johnson-Laird, Khemlani, & Goodwin, 2015). In the mental model framework, the disjunction of p and q is represented with multiple models, or simulations, of alternative possibilities: "p," "q" (assuming p and q are mutually exclusive). The disjunctive inference is carried out with an algorithm that erases the alternatives incompatible with new data. Although such an algorithm carries out deductively-valid inferences, it involves no discrete logical operators (e.g., a collection of models is not a representation that can be recursively combined with "NOT p").

Crucially, without logical operators, disjunction requires to store and update multiple mental models in parallel. This is costly for adults and plausibly very challenging for infants with immature cognitive resources (Gauffroy & Barrouillet, 2011). Thus, we may expect that if infants have no logical operators, they will simulate and store just one disjunct at the time (for a related

prediction, see Leahy & Carey, 2020). A new study provides evidence that infants do NOT respond to objects with multiple possible identities by simulating just a single identity at the time (Cesana-Arlotti, Varga, & Téglás, 2022), as their pupil diameter – indexing processing load – increases with the number of possible identities. Although infants might have simulated multiple models, the proposal of a single complex logical representation (e.g., “the elephant OR the ball”) may best account for this result, considering their limited cognitive resources. Future research will further test the constituency structure of preverbal disjunctive representations.

In conclusion, the reemergence of LoTH is a boon for developmental psychologists, logicians, and philosophers alike: It points to the need to chart the foundation of our logical capacities and opens exciting questions about the logical primitives of the mind.

Acknowledgments. I am grateful to E. Téglás, F. Keil, and L. Barlassina for their very valuable comments on the manuscript.


Financial support. This research was supported by funds for a postdoctoral fellowship from the James S. McDonnell Foundation.

Competing interest. None.

References

- Bermúdez, J. L. (2007). *Thinking without words* (1st issued as an Oxford University Press paperback). Oxford University Press.
- Braine, M. D. S., & O'Brien, D. P. (1998). *Mental logic*. Erlbaum.
- Burge, T. (2010). Steps toward origins of propositional thought. *Disputatio*, 4(29), 39–67. <https://doi.org/10.2478/disp-2010-0010>
- Cesana-Arlotti, N., Kovács, Á. M., & Téglás, E. (2020). Infants recruit logic to learn about the social world. *Nature Communications*, 11(1), 5999. <https://doi.org/10.1038/s41467-020-19734-5>
- Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science (New York, N.Y.)*, 359(6381), 1263–1266. <https://doi.org/10.1126/science.1263359>
- Cesana-Arlotti, N., Varga, B., & Téglás, E. (2022). The pupillometry of the possible: An investigation of infants' representation of alternative possibilities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1866), 20210343. <https://doi.org/10.1098/rstb.2021.0343>
- Chierchia, G. (2013). *Logic in grammar: Polarity, free choice, and intervention* (1st ed.). Oxford University Press.
- Fodor, J. A. (1979). *The language of thought* (1st paperback printing). Harvard University Press.
- Gauffroy, C., & Barrouillet, P. (2011). The primacy of thinking about possibilities in the development of reasoning. *Developmental Psychology*, 47(4), 1000–1011. <https://doi.org/10.1037/a0023269>
- Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1), 108–154. <https://doi.org/10.1080/03640210701802071>
- Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Laurence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 623–654). MIT Press.
- Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, 19(4), 201–214. <https://doi.org/10.1016/j.tics.2015.02.006>
- Leahy, B. P., & Carey, S. E. (2020). The acquisition of modal concepts. *Trends in Cognitive Sciences*, 24(1), 65–78. <https://doi.org/10.1016/j.tics.2019.11.004>
- Luo, Y., Hennefeld, L., Mou, Y., vanMarle, K., & Markson, L. (2017). Infants' understanding of preferences when agents make inconsistent choices. *Infancy*, 22(6), 843–856. <https://doi.org/10.1111/infa.12194>
- Piantadosi, S. T. (2021). The computational origin of representation. *Minds and Machines*, 31(1), 1–58. <https://doi.org/10.1007/s11023-020-09540-9>
- Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, 123(4), 392–424. <https://doi.org/10.1037/a0039980>
- Pietroski, P. M. (2018). *Conjoining meanings: Semantics without truth values* (1st ed.). Oxford University Press.
- Quilty-Dunn, J., & Mandelbaum, E. (2018). Against dispositionalism: Belief in cognitive science. *Philosophical Studies*, 175(9), 2353–2372. <https://doi.org/10.1007/s11098-017-0962-x>

The computational and the representational language-of-thought hypotheses

David J. Chalmers 

Department of Philosophy, New York University, New York, NY, USA
chalmers@nyu.edu; consc.net/chalmers

doi:10.1017/S0140525X23001796, e269

Abstract

There are two versions of the language-of-thought hypothesis (LOT): Representational LOT (roughly, structured representation), introduced by Ockham, and computational LOT (roughly, symbolic computation) introduced by Fodor. Like many others, I oppose the latter but not the former. Quilty-Dunn et al. defend representational LOT, but they do not defend the strong computational LOT thesis central to the classical-connectionist debate.

There are two versions of the language-of-thought hypothesis. The representational language-of-thought hypothesis (r-LOT), introduced by William of Ockham and defended by Quilty-Dunn et al., concerns the structure of mental representation. The computational language-of-thought hypothesis (c-LOT), introduced by Jerry Fodor, concerns computation over mental representations. r-LOT is much weaker than c-LOT and is more widely accepted. I accept the former but reject the latter. As a result, I agree with many of Quilty-Dunn et al.'s conclusions while finding that they have not really defended the most controversial form of LOT.

In more detail: r-LOT (I use “LOT” for both the language and the hypothesis) says roughly that thought involves sententially structured mental representations. At minimum, there are nominal representations (e.g., *Biden*) and predicative representations (e.g., *president*) that combine into structured representations (e.g., *Biden is president*) with propositional content. Structured representations may also involve connectives (e.g., *and*), quantifiers (e.g., *all*), operators (e.g., *always*), and other types familiar from the linguistic case.

r-LOT is not trivially true, but it is plausible and hard to deny. It follows naturally from the claims that (1) people make judgments such as *Biden is president*, (2) these judgments involve combining nominal and predicative representations (or concepts, in the sense where concepts are mental representations) such as *Biden* and *president*, and (3) these representations can be recombined in judgments such as *Biden is from Delaware*. My sense is that most contemporary cognitive scientists and philosophers of mind accept these fairly weak claims. Importantly, these claims do not have immediate consequences regarding computation or cognitive architecture.

The c-LOT adds to r-LOT the key claim that thought involves computation over these sententially structured representations. The classical version of this hypothesis says that r-LOT representations are the medium through which all cognitive computation takes place. That is, the basic vehicles of representation in the r-LOT system (atomic words in the representational language-of-thought) also serve as the basic vehicles of computation

(atomic computational states to which cognitive algorithms apply).

The c-LOT hypothesis was canonically formulated by Jerry Fodor's book *The Language of Thought* (1975). Computation plays a central role throughout the book, from the main argument for LOT at the start of chapter 1 ("Computation presupposes a medium of computation: a representational system," p. 27) to the conclusion ("More exactly: Mental states are relations between organisms and internal representations, and causally interrelated mental states succeed one another according to computational principles which apply formally to the representations," p. 198). There are other works (e.g., "Propositional Attitudes") in which Fodor focuses mainly on r-LOT, but computation is central in the canonical statement.

(Related distinctions: Fodor himself [1980] distinguishes the "representational theory of mind" and the "computational theory of mind" [though neither requires a language-of-thought]. Rescorla [2017] distinguishes a core version of LOT that involves "representational theory of thought" plus "compositionality of thought" and perhaps "logical structure" [in my terms, a version of r-LOT] from a stronger version that adds "the classical computational theory of mind" [yielding a version of c-LOT, though I understand the computational constraint differently from Rescorla].)

Most work in symbolic artificial intelligence (AI) uses a version of c-LOT. Both involve computation over atomic symbols: Entities that are both representationally atomic and computationally atomic. Atomic symbols have no computationally relevant internal structure (if they did, they would not be computationally atomic). Instead, their internal form is arbitrary.

The most significant opposition to LOT, in the classical-connectionist debate, has been opposition to c-LOT. In most neural network models there are no computationally atomic symbols. Representations are distributed over multiple quasi-neural units. As a result, in these models computation is *subsymbolic computation*: Computation takes place among units below the level of representation. Because computational primitives (units) are not representational primitives in these models, representation is not the medium of computation. Subsymbolic computation is incompatible with c-LOT.

At the same time, subsymbolic computation is quite compatible with r-LOT. This is clearest in the work of structured connectionists (e.g., Chalmers, 1990; Smolensky, 1988), where distributed representations (e.g., of *Biden* and *president*) can combine with each other systematically to yield new distributed representations such as *Biden is president*. This is naturally seen as a structured representational system involving subsymbolic computation: r-LOT without c-LOT. The structured connectionist research program is still a work in progress, but it is arguable that contemporary large language models also combine structured representation (of facts such as *Biden is president*) with subsymbolic computation. A second and third way of combining r-LOT with subsymbolic computation are provided by the framework of vector symbolic architectures (Kleyko et al., 2022), where representations are vectors, and Piantadosi's combinator framework (2021), where the computational primitives S and K fall below the level of representation.

(Terminology: All three of these are computational versions of r-LOT in a broad sense. In an alternative phraseology, one might call the Fodorian version the *classical* computational LOT (cc-LOT), while calling subsymbolic versions *nonclassical* computational LOT (nc-LOT). But I will reserve "c-LOT" for the classical Fodorian version.)

Proponents of LOT often argue that structured connectionism is merely an implementation of LOT. We can now see that this claim is false or at best misleading. Implementation is standardly a computational relation between algorithms, requiring the implementing algorithm to be a more fine-grained version of the implemented algorithm with the same input/output behavior. The most interesting subsymbolic algorithms (e.g., in artificial neural networks) are never implementations of symbolic algorithms in this sense. The success of the deep-learning paradigm has provided strong evidence that the behavior of these systems (especially their success in learning and generalizing, but also their post-learning success) is not the result of implementing a more coarse-grained symbolic algorithm and cannot be duplicated by such algorithms. These systems may realize an r-LOT, but they do not implement a c-LOT. The quasi-symbolic operations of composition, decomposition, and quasi-logical inference may be available, but they are a tiny subset of the operations one can perform on the relevant distributed representations. As I argued in Chalmers (1990), one can also perform all sorts of holistic operations on distributed representations that do not proceed via these symbolic operations. It is plausibly subsymbolic operations like this that are largely responsible for the remarkable capacities of neural network systems.

Quilty-Dunn et al. don't make the distinction between r-LOT and c-LOT in their article, but their LOT appears to be a version of r-LOT. Their six core claims defining LOT do not mention computation (except in one case, incidentally). Four of the key claims (role-filler independence, predicate-argument structure, logical operators, abstract conceptual content) clearly pertain to representation but not computation. A fifth (inferential promiscuity) mentions computational theories of logical inference as versions of LOT, but computation does not play a defining role, and inferential promiscuity can equally be present in r-LOT without c-LOT (e.g., Ockham-style or subsymbolic systems).

The requirement of "discrete constituents" may suggest c-LOT, though it doesn't mention computation explicitly. Distributed representations in a structured connectionist systems arguably aren't discrete in the authors sense, in that representation of *Biden* and of *president* (say) can be intertwined nondiscretely in a representation of *Biden is president*. On the contrary, many subsymbolic computational systems involve discrete constituents without c-LOT. Piantadosi's system is one. Another is provided by the word embedding format for representing words that is ubiquitous in current language models. Here words are represented by multidimensional vectors where individual units often lack any clear semantic significance. "Biden is president" may be represented as a sequence of vectors for the individual words, so the constituents are discrete, but representations remain distributed and processing remains subsymbolic. So the discrete representational constituents do not require c-LOT.

Now, perhaps the absence of a computational constraint is an easily correctable omission. Quilty-Dunn et al. discuss computational approaches at some length in other sections of their article. They could easily enough add a seventh constraint connecting computation to representation, holding that the representational primitives are computationally primitive and serve as the medium of computation. The trouble is that strong evidence for this seventh claim is much harder to find.

The target article does argue that many Bayesian theorists provide computational accounts involving a "probabilistic LOT"

associated with sententially structured representations. This suggests r-LOT, but it does not obviously lead to c-LOT, as Bayesian accounts are usually not cast at the algorithmic level (rather, at Marr's higher "computational" level). These accounts have many algorithmic implementations, including subsymbolic implementations in deep-learning systems. So there is no obvious strong evidence for c-LOT here, and any evidence would need to be stacked against the counterevidence provided by deep-learning models.

Overall: If Quilty-Dunn et al. are defending c-LOT, then more work is needed to make the defense explicit. If they are defending only r-LOT, then their conclusion is plausible, and my only objection is one of relative unambition.


Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

- Chalmers, D. J. (1990). Syntactic transformations on distributed representations. *Connection Science*, 2(1), 53–62.
- Fodor, J. A. (1975). *The language of thought*. Harvard University Press.
- Fodor, J. A. (1980). *Representations: Philosophical essays on the foundations of cognitive science*. MIT Press.
- Kleyko, D., Davies, M., Frady, E. P., Kanerva, P., Kent, S. J., Olshausen, B. A., ... Rabaey, J. M. (2022). Vector symbolic architectures as a computing framework for emerging hardware. *Proceedings of the IEEE*, 110, 1538.
- Piantadosi, S. T. (2021). The computational origin of representation. *Minds and Machines*, 31, 1–58.
- Rescorla, M. (2017). The language of thought hypothesis. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.
- Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1), 1–23.

The language of tactile thought

Tony Cheng 

Department of Philosophy / Research Center for Mind, Brain and Learning,
National Chengchi University, Taipei, Taiwan

h.cheng.12@ucl.ac.uk
www.tonycheng.net

doi:10.1017/S0140525X2300208X, e270

Abstract

The target article argues that language-of-thought hypothesis (LoTH) is applicable to various domains, including perception. However, it focusses exclusively on the visual case, which is limited in this regard. I argue for two ideas in this commentary: first, their case can be extended to other modalities such as touch; and second, the status of those six criteria needs to be further clarified.

In the target article, Quilty-Dunn et al. not only revive the language-of-thought hypothesis (LoTH) in the original domain, but also extend it to new domains such as perception, where the hypothesis did not cover in the past (Fodor, 1975, 2008; Schneider, 2011), for obvious reason: Even if LoTH can be made plausible enough for propositional thoughts, it is unlikely to make it work for imagistic contents, such as those in

perceptions (e.g., Block, 2007; Campbell, 1997; Fodor, 2007). The authors, however, argue that “[i]f cognition is largely LoT-like, and perception feeds information to cognition, then we should expect at least some elements of perception to be LoT-like” (target article, sect. 4, para. 2). More specifically, they invoke six criteria to make the case that some elements in perception such as object files and structured relations are LoT-like. Although their case here is indeed strong, their examples are exclusively visual, and therefore limited in this regard. In what follows I will extend their proposal by providing examples from touch that also exemplify some LoT-like properties, though with the proviso that there are some grey areas to be concerned about.

The relevant examples are the so-called “tactile field” cases in the recent empirical and philosophical literatures (Green, 2022; Skrzypulec, 2021, 2022). In those cases, multiple tactile stimuli constitute spatial patterns that facilitate varieties of tactile judgements (Cheng, 2019, 2020, 2022; Fardo, Beck, Cheng, & Haggard, 2018; Haggard & Giovagnoli, 2011). Although in those cases researchers often emphasise the *holistic* characters of tactile pattern perceptions, the tactile fields also have some LoT-like structures identified by the authors, and this might strengthen their case that LoTH can extend to perceptions, including nonvisual ones. Let's look into some relevant details.

The tactile field cases typically involve multiple tactile stimuli, each of them exists independent of one another. This exemplifies *discrete constituents* (property 1). Those tactile stimuli can exhibit different properties at different times; for example, some of them can vibrate while the other ones remain still. This exemplifies *predicate–argument structure* (property 3). Moreover, the multiple tactile stimuli can jointly vibrate to generate geometrical representations such as lines, triangles, and squares; as long as the stimuli in question generate neutral touch (as opposed to thermally salient or nociceptive feels), those shapes can be equally represented. This exemplifies *abstract conceptual content* (property 6). Thus, tactile fields at least exemplify three core properties of LoTs.

One might argue that there are six core properties identified by Quilty-Dunn et al., but tactile fields might exemplify only three of them. Is this enough? Well, in their discussion of object files in section 4.1, they also primarily point out that properties 1, 3, and 6 are exemplified by them. Moreover, later they write that “perceptual representations of individual objects contain discrete constituents that are organized in a predicate–argument structure and predicate abstract conceptual contents” (target article, sect. 4.2, para. 1), and these are again properties 1, 3, and 6. It seems that amongst the six core properties, somehow these three are more important, or even almost definitive.

Now, the good news is that the tactile field cases are at least as good as the object file cases, so if the latter fits LoTH, so does the former. However, this generates a potential worry about the status of the six core properties. It should be quite clear that Quilty-Dunn et al. are not engaging the traditional project of offering necessary and sufficient conditions for LoT: They write that “[m]any, perhaps all, of these properties are not necessary for a representational scheme to count as an LoT.... We regard these properties as (somewhat) independent axes on which a format can be assessed for how LoT-like it is” (target article, sect. 2, para. 3). But isn't this too weak? If none of them is necessary, and presumably none of them is by itself sufficient, how do we assess whether a given format fits LoTH? Relatedly, to say “LoT-like” might make the situation worse, as similarities are too vague to

be useful without further explications. The positive story offered by Quilty-Dunn et al. invokes the notion of “cluster,” and they write that “LoTH predicts that these sorts of evidence should tend to cooccur” (target article, sect. 2, para. 13). For them, such “clustering-based approach” provides “an abductive, empirical argument for LoTH” (target article, sect. 2, para. 13). There is nothing wrong with this approach as such, but one is justified in asking for more concrete criteria: Do these six core properties constitute a weighting system? Are properties 1, 3, and 6 indeed more important than the others? It is even more worrying that Quilty-Dunn et al. use the term “core” to name these six properties, as that signifies that even if they are not *necessary*, they might be near enough.

As mentioned above, Quilty-Dunn et al. write that “[i]f cognition is largely LoT-like, and perception feeds information to cognition, then we should expect at least some elements of perception to be LoT-like, because the two systems need to interface.” This general point is well taken, but again, it might be too vague to be truly useful. Consider a similar thought in the conceptualism debate: If cognition is largely conceptual, and perception feeds information to cognition, then we should expect at least some elements of perception to be conceptual, because the two systems need to interface. Here we face two worries: First, “largely” is unclear, and second, if only *some* elements are conceptual, what about those nonconceptual elements? How do they feed information to cognition (McDowell, 1996)? If these are legitimate challenges to partial conceptualism (Peacocke, 1992), then the same doubt can be cast on the general point made by Quilty-Dunn et al.

Let’s head back to the good news. In the target article, Quilty-Dunn et al. provide many rationales for the thesis that LoTH is the best game in town. They might well be right about that. The tactile field cases, and potential cases from other sensory modalities, should strengthen their hypothesis. That said, it will be very helpful if the status of the six core properties can be further clarified.

Acknowledgments. I would like to thank all of my past teachers and colleagues in New York and London for helping me think through these issues in the past few years.

Financial support. This study was supported by the Ministry of Science and Technology, Taiwan (MOST 109-2410-H-004-006-MY3).

Competing interest. None.

References

- Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Science*, 30(5), 481–548.
- Campbell, J. (1997). Sense, reference, and selective attention. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 71, 55–74.
- Cheng, T. (2019). On the very idea of a tactile field, or: A plea for skin space. In T. Cheng, O. Deroy, & C. Spence (Eds.), *Spatial senses: Philosophy of perception in an age of science* (pp. 226–247). Routledge.
- Cheng, T. (2020). Molyneux’s question and somatosensory spaces. In G. Ferretti & B. Glenney (Eds.), *Molyneux’s question and the history of philosophy* (pp. 300–312). Routledge.
- Cheng, T. (2022). Spatial representations in sensory modalities. *Mind and Language*, 37(3), 485–500.
- Fardo, F., Beck, B., Cheng, T., & Haggard, P. (2018). A mechanism for spatial perception on human skin. *Cognition*, 178, 236–243.
- Fodor, J. A. (1975). *The language of thought*. Harvard University Press.
- Fodor, J. A. (2007). The revenge of the given. In B. P. McLaughlin & J. D. Cohen (Eds.), *Contemporary debates in philosophy of mind*. Blackwell.
- Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. Oxford University Press.
- Green, E. J. (2022). Representing shape in sight and touch. *Mind and Language*, 37(4), 694–714.

Haggard, P., & Giovagnoli, G. (2011). Spatial patterns in tactile perception: Is there a tactile field? *Acta Psychologica*, 137(1), 65–75.

McDowell, J. (1996). *Mind and world*. Harvard University Press.

Peacocke, C. (1992). *A study of concepts*. MIT Press.

Schneider, S. (2011). *The language of thought: A new philosophical direction*. MIT Press.

Skrzypulec, B. (2021). Spatial content of painful sensations. *Mind and Language*, 36(4), 554–569.

Skrzypulec, B. (2022). Is there a tactile field? *Philosophical Psychology*, 35(3), 301–326.

Concept learning in a probabilistic language-of-thought. How is it possible and what does it presuppose?

Matteo Colombo 

Tilburg Center for Logic and Philosophy of Science (TiLPS), Tilburg University, Tilburg, The Netherlands

m.colombo@uvt.nl; <https://mteocolphi.wordpress.com/>

doi:10.1017/S0140525X23002029, e271

Abstract

Where does a probabilistic language-of-thought (PLoT) come from? How can we learn new concepts based on probabilistic inferences operating on a PLoT? Here, I explore these questions, sketching a traditional circularity objection to LoT and canvassing various approaches to addressing it. I conclude that PLoT-based cognitive architectures can support genuine concept learning; but, currently, it is unclear that they enjoy more explanatory breadth in relation to concept learning than alternative architectures that do not posit any LoT.

Quilty-Dunn et al. survey empirical evidence consistent with Bayesian models of cognition to advertise the explanatory breadth of language-of-thought (LoT)-based cognitive architectures. They show that Bayesian models that treat *concepts* as stochastic programmes and *thinking* as approximate Bayesian inference can fit various sets of experimental data. But they do not say much about the origin of the representational system constituting a probabilistic LoT (PLoT), the nature of learning supported by LoT-based architectures implementing probabilistic inference, and how the explanatory breadth of such architectures is more convincing compared to architectures that do not posit any LoT.

(In)famously, Jerry Fodor developed a circularity objection to LoT. If concept learning is a process of inductive inference aimed at testing hypotheses concerning the identity conditions for a given concept, then this process must recruit the very concept it seeks to learn. But if that is the case, then no new concept can be learned through inductive inference aimed at hypothesis testing (2008, p. 139). Furthermore, if the very representational system constituting an LoT cannot be learned through inductive inference – because that would also generate a vicious circularity – then all representations constituting an LoT must be innate (1975, p. 65).

Similar objections apply to PLoT-based architectures too, where the problem of concept learning can be understood as the problem of performing Bayesian inference to compute the

posterior probability of hypotheses consisting in stochastic programmes formulated in a PLoT, given, say, a set of observed example objects and observed labels for those objects (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Goodman, Tenenbaum, & Gerstenberg, 2015). Because the hypothesis space of stochastic programmes in PLoT-based architectures is prespecified, these architectures do not seem to support genuine learning of any new concept. The probabilistic inferences that they implement are aimed at updating and comparing the posterior probabilities of hypotheses that they possess from the outset, which implies an implausibly strong nativist picture of cognitive development (see, e.g., Elman et al., 1996; Putnam, 1988, Ch. 1).

To address this circularity objection, one approach is to distinguish between *learning* and *acquiring* a concept, and point out that thinkers acquire concepts without learning them – where *learning* consists in a rationally evaluable process, whereas *acquiring* a concept is a nonrational, purely causal process, driven, for example, by associative processes implemented in a connectionist architecture (Fodor, 1975, 2008). But if nonrational, noncognitive, associative processes provide us with the best explanation of concept acquisition in many domains, then PLoT-based architectures would enjoy significantly less explanatory breadth than what Quilty-Dunn et al. suggest.

A different approach is to insist that thinkers learn concepts in a way that is rationally evaluable and is based on probabilistic inferences in PLoT-based architectures, and to also emphasize that this learning process does not need to generate any vicious circularity, or be committed to an implausible nativism.

Carey (2009), for example, argues that children learn concepts based on Quinian-bootstrapping processes operating on innate, core systems of knowledge. Although the notions of “bootstrapping” and “core knowledge” have been helpful for developing empirically adequate PLoT-based models (e.g., Piantadosi, Tenenbaum, & Goodman, 2012), one objection is that bootstrapping in a PLoT-based architecture cannot really explain concept learning, because its built-in knowledge would already include the very concepts that it purports to explain (cf. Beck, 2017, for a critical assessment of this objection). In this conceptualization, learning would amount to combining and recombining built-in representations based on built-in rules of composition. But then, one may wonder how combining and recombining a stock of built-in representations constituting one’s core knowledge qualifies as genuine learning, and, more substantially why building domain-specific, core knowledge into a PLoT-based architecture is not a mere exercise in *ad hoc* modelling.

To resolve these issues, we should notice that, first, any conception of learning without any built-in hypothesis space is incoherent; second, a learner’s hypothesis space is hierarchically organized, and includes stacks of *latent* and *explicit* hypothesis spaces (Perfors, 2012); third, in PLoT-based architectures, there is ample latitude for the choice of built-in hypotheses/representations and learning rules (Colombo, 2019). But this choice – though it is often left unconstrained by evolutionary, neurobiological, and psychological evidence – is typically transparent and empirically evaluable, which facilitates clearer understanding of the nativist (or empiricist) character of any given PLoT-based architecture compared to connectionist ones (Colombo, 2018).

Considering these three points, it is easier to appreciate why we should reject the worries that PLoT-based architectures must presuppose an unacceptable amount of innate structure and cannot support genuine learning. If a learner’s *latent* hypothesis space defines the learner’s representational capacity – that is,

the range and kinds of possible thoughts that the learner can entertain over a lifetime – then some latent hypothesis space defined by abstract primitives (sort of Kantian categories) is built into any PLoT-based architecture. Manipulating these abstract primitives can generate a learner’s *explicit* hypothesis space, which defines the learner’s actual thoughts at a given time. Such thoughts can play various causal roles in perception, action, and other cognitive functions, but are *not* built into the learner: They are generated from the latent hypothesis space. As Perfors (2012, pp. 131–132) helpfully puts it, a latent hypothesis space is like a typewriter with an infinite amount of paper, which can generate certain kinds of documents like *Paradise Lost*, but not others like *La Madonna della Pietà*; the set of actual documents that have been typed out and can enter various causal relationships (e.g., *Paradise Lost* can be read or burnt) is like an explicit hypothesis space. Given this distinction, concept learning consists in an extended, hierarchically organized process of hypothesis generation and hypothesis testing tapping abstract primitives defining the learner’s overall representational power.

But although this distinction helps us to successfully address a traditional theoretical objection to LoT, showing that PLoT-based architectures can support genuine concept learning without necessarily positing an implausible amount of innate structure, it remains an open empirical question whether PLoT-based architectures enjoy more explanatory breadth in relation to concept learning compared to architectures that do not posit any LoT.

Competing interest. None.

References

- Beck, J. (2017). Can bootstrapping explain concept learning?. *Cognition*, 158, 110–121.
- Carey, S. (2009). *The origin of concepts*. Oxford University Press.
- Colombo, M. (2018). Bayesian cognitive science, predictive brains, and the nativism debate. *Synthese*, 195, 4817–4838.
- Colombo, M. (2019). Learning and reasoning. In M. Sprevak & M. Colombo (Eds.), *The Routledge handbook of the computational mind* (pp. 381–396). Routledge.
- Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. MIT Press.
- Fodor, J. (1975). *The language of thought*. Harvard University Press.
- Fodor, J. (2008). *LOT 2: The language of thought revisited*. Oxford University Press.
- Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32, 108–154.
- Goodman, N., Tenenbaum, J., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Laurence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 623–654). MIT Press.
- Perfors, A. (2012). Bayesian models of cognition: What’s built in after all? *Philosophy Compass*, 7(2), 127–138.
- Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition* 123(2), 199–217.
- Putnam, H. (1988). *Representation and reality*. MIT Press.

Putting relating at the core of language-of-thought

Jan De Houwer 

Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium
jan.dehouwer@ugent.be
www.liplab.be

doi:10.1017/S0140525X23001978, e272

Abstract

Propositional representations are units of information with a relational content. Their relational nature allows for the six distinctive properties of language-of-thought representations. Putting relating at the core of language-of-thought also fits well with the idea that thinking and reasoning are instances of relational behavior. These propositional and behavioral perspectives can be combined within a functional-cognitive framework.

I agree with Quilty-Dunn et al. that, from a cognitive point of view, thinking in human and nonhuman organisms relies on language-like structured representations. In my own work, I have referred to these representations as propositional representations. For many years now (e.g., Boddez, De Houwer, & Beckers, 2017; De Houwer, 2009, 2014), my colleagues and I have argued that seemingly simple phenomena such as conditioning, implicit evaluation, and habitual responding are mediated by this type of representations (see De Houwer, 2019, for a review). In line with Quilty-Dunn et al., we pointed out that propositional representations do not necessarily have the same structure as natural language and therefore can be present also in nonverbal organisms (De Houwer, Hughes, & Barnes-Holmes, 2016). Rather than focusing on the many communalities between our views, in this commentary, I highlight a few differences so as to further stimulate the scientific debate on the nature of thought.

Whereas Quilty-Dunn et al. put forward six distinctive properties of “language-of-thought” representations, I have characterized propositional representations in terms of one core property: Their relational nature (e.g., De Houwer, 2018; also see Lagnado, Waldmann, Hagmayer, & Sloman, 2007). More specifically, a propositional representation can be defined as a unit of information with a relational content. In principle, this information can be implemented in many physical vehicles (e.g., a brain, an artificial associative network) but it needs to specify the way in which elements in the world are related (e.g., element A “is a,” “has a,” “belongs to,” “causes,” “predicts,” ... element B). In my opinion, the properties put forward by Quilty-Dunn et al. are implied by this one core property: Relating requires discrete constituents (e.g., elements A and B), requires role-filler independence (e.g., whether A is the cause or the effect of B), is truth-evaluable (e.g., to evaluate whether A is a cause of B), allows for logical operators (e.g., A AND B causes C), allows for inferential promiscuity (e.g., to infer that B will follow A), and allows for abstract conceptual content (e.g., the concept of causality). It would be interesting to know whether Quilty-Dunn et al. see any reason for not putting relating at the core of language-of-thought representations.

A second way in which my work deviates from that of Quilty-Dunn et al. is that I adopt a functional-cognitive framework in which psychological phenomena are conceived of in behavioral terms (De Houwer, 2011; Hughes, De Houwer, & Barnes-Holmes, 2016a). From this perspective, psychological phenomena can be mediated by propositional representations but can also be studied without referring to any type of representation. Although Quilty-Dunn et al. refer to Skinner’s behaviorism as a relic, my colleagues and I see much merit in the work of Skinner and those inspired by Skinner. In particular, we have linked our propositional theories to relational frame theory (RFT), which builds on the work of Skinner but goes beyond this work by postulating the concept of arbitrarily applicable relational responding (AARR; Hayes, Barnes-Holmes, & Roche,

2001). Relational responding is responding to one stimulus in terms of another stimulus. It can be grounded in nonarbitrary features (e.g., physical features or direct training with those features) as is the case when a rat presses a lever for food as a function of the relative length of lines (e.g., if a blue line is longer than a red line). Humans, however, can also respond relationally in arbitrarily applicable ways (i.e., not grounded in physical features or direct training with those features). For instance, they can select a dime as being more than a nickel in terms of monetary value even though a dime is less than a nickel in terms of size.

The ideas of behavioral researchers like Skinner (1953) and Hayes et al. (2001) played a vital role in our research on conditioning, implicit evaluation, and habitual responding. When my colleagues and I started this research, these phenomena were often defined in terms of associative representations (e.g., conditioning as the formation of associations in memory). By adhering to behavioral definitions of those phenomena (e.g., conditioning as the impact of stimulus pairings on behavior), we could at least raise the possibility that these phenomena are mediated by propositional representations (see De Houwer, 2019; De Houwer, Van Dessel, & Moran, 2021). Moreover, it allowed us to link those phenomena with the literature on AARR (e.g., De Houwer, Finn, Raemaekers, Cummins, & Boddez, *in press*; Hughes, De Houwer, & Perugini, 2016b).

In line with the ideas of Skinner (1953) and Hayes et al. (2001), I believe that there is merit in adopting a behavioral perspective on thinking and reasoning in general. It would imply that thinking and reasoning, like other behaviors, are a function of their antecedents and consequences (see De Houwer, 2022, for a discussion). From the perspective of RFT, thinking and reasoning are covert forms of one specific type of behavior: AARR. Because of its emphasis on relational responding, a behavioral RFT perspective on thinking and reasoning is highly compatible with the cognitive idea that thinking and reasoning rely on propositional (i.e., relational) representations (also see McLoughlin, Tyndall, & Pereira, 2020). The added value of adopting this behavioral perspective on thinking and reasoning is that it (a) offers a new way of talking about thinking and reasoning that is abstract, precise, and separated from folk psychology terms, (b) sheds new light on the difference in thinking and reasoning in verbal and nonverbal organisms (De Houwer et al., 2016), (c) allows researchers to relate knowledge about the moderators of AARR to knowledge about thinking and reasoning, which (d) includes ideas about how thinking and reasoning is shaped during the learning history of organisms (and therefore how developmental deficits in thinking and reasoning can be remedied; De Houwer et al., *in press*). I therefore hope that cognitive scientists will explore and exploit what a behavioral perspective on thinking and reasoning has to offer.

Financial support. The preparation of this paper was made possible by Ghent University Grant BOF22/MET_V/002 to Jan De Houwer.

Competing interest. None.

References

- Boddez, Y., De Houwer, J., & Beckers, T. (2017). The inferential reasoning theory of causal learning: Towards a multi-process propositional account. In M. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 1–22). Oxford University Press.
- De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior*, 37, 1–20.
- De Houwer, J. (2011). Why the cognitive approach in psychology would profit from a functional approach and vice versa. *Perspectives on Psychological Science*, 6, 202–209.

- De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, 8, 342–353.
- De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin*, 13(3), Article e28046. <https://doi.org/10.5964/spb.v13i3.28046>
- De Houwer, J. (2019). Moving beyond the distinction between system 1 and system 2: Conditioning, implicit evaluation, and habitual responding might also be mediated by relational knowledge. *Experimental Psychology*, 66, 257–265.
- De Houwer, J. (2022). On the merits and challenges of treating conscious and unconscious thoughts and feelings as behavior. *PsychArchives*. <https://doi.org/10.23668/psycharchives.5332>
- De Houwer, J., Finn, M., Raemaekers, M., Cummins, J., & Boddez, Y. (in press). Thinking of learning phenomena as instances of relational behavior. *Learning & Behavior*.
- De Houwer, J., Hughes, S., & Barnes-Holmes, D. (2016). Associative learning as higher-order cognition: Learning in human and nonhuman animals from the perspective of propositional theories and relational frame theory. *Journal of Comparative Psychology*, 130, 215–225.
- De Houwer, J., Van Dessel, P., & Moran, T. (2021). Attitudes as propositional representations. *Trends in Cognitive Sciences*, 25, 870–882.
- Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. Kluwer.
- Hughes, S., De Houwer, J., & Barnes-Holmes, D. (2016a). The moderating impact of distal regularities on the effect of stimulus pairings: A novel perspective on evaluative conditioning. *Experimental Psychology*, 63, 20–44.
- Hughes, S., De Houwer, J., & Perugini, M. (2016b). The functional-cognitive framework for psychological research: Controversies and resolutions. *International Journal of Psychology*, 51, 4–14.
- Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). Oxford University Press.
- McLoughlin, S., Tyndall, I., & Pereira, A. (2020). Convergence of multiple fields on a relational reasoning approach to cognition. *Intelligence*, 83, 101491.
- Skinner, B. F. (1953). *Science and human behavior*. Macmillan.

Developmental and multiple languages-of-thought

Andreas Demetriou^{a,b,c} 

^aCyprus Academy of Sciences, Letters, and Arts, Nicosia, Cyprus; ^bUniversity of Cyprus, Nicosia, Cyprus and ^cUniversity of Nicosia, Nicosia, Cyprus
ademetriou@ucy.ac.cy

doi:10.1017/S0140525X23002005, e273

Abstract

We agree with the target article that assuming language-of-thought (LoT) is useful for the development of cognitive and developmental theories. We note that the target article is weak in its assumptions about development of LoT and possible existence of multiple LoTs. In response to these weaknesses, we outline several developmental principles for LoT development, showing how a developmental theory of LoT springs from probabilistic LoT. We suggest a system 1.5 of reasoning allowing interchange between Bayesian and logical rules as it fits purposes or domain.

1. Introduction

This commentary focuses on section 5 of the target article, dealing with language-of-thought (LoT) in children. We agree with the target article that assuming LoT is useful for further development of cognitive and developmental theories. Specifying constructs in LoT offers a system for exploring relations between representations and their development. The six properties of LoT proposed in the target article are useful, enabling to specify

mental units carrying information, rules for binding and transforming units, and model the development of mind. Examples of the developmental implementation of properties are given below. However, the target article is weak in two themes: Development and multiplicity of LoT.

2. Development of LoT

The target theory is minimally developmental. Being limited in infancy, it avoids to account for changes in LoT or specifies how rules, principles, and constraints of LoT change in concern to LoT properties. We outline some general principles of a developmental theory of LoT (DLoT) as complementary to the target theory. DLoT claims that a probabilistic language-of-thought (PLoT), defended in the target article, provides early foundations of LoT, but it does not accommodate later development. DLoT argues that rules emerge from PLoT with development, upgrading pragmatic reasoning in early childhood into deductive and analogical reasoning later. Mental awareness is a critical factor in this development (Demetriou, Makris, Kazi, Spanoudis, & Shayer, 2018). In psychometric terms, DLoT is a systematic expansion of a core relational integration capacity into representations and rules prescribing optimal inference-based integration to handle novel encounters capitalizing on experience (Demetriou, Golino, Spanoudis, Maris, & Greiff, 2021).

This is obvious in mastering the four basic schemes of syllogistic reasoning: Modus ponens (MP), modus tollens (MT), and the fallacies, affirming the consequent (AC) and denying the antecedent (DA). MP, a logical primitive, emerges as a Bayesian product from pragmatic contexts, at 5–6 years of age (Oaksford & Chater, 2020). Pragmatic MP is an induction binding two representations (“A occurs” and “B occurs”) into an inductive rule (i.e., “When A occurs, B also occurs”). Transition to rule-based thinking at 6–7 years lifts the alignments of representations of preschool age into a rule-based representational imperative (A and B, A, therefore B), involving a sense of logical necessity. MT is grasped at 8–9 years, when inferential rules emerge as explicit constraints on how representations may be combined in syllogistic chains. By the end of childhood, representational imperatives are fluent enough to be read both ways (A and B, not B, therefore not A) (Christoforides, Spanoudis, & Demetriou, 2016).

The integration of MP and MT into a fluent inferential ensemble transforms inductive imperatives into deductive necessities constrained by rules explicitly metarepresented in principles specifying how inferential spaces are interrelated. Rules specify that different representational spaces may have different inferential constraints (e.g., birds fly, mammals walk, fish swim, etc.) yielding different inductive implications about individual elements in each space (e.g., blackbirds fly, elephants walk, sharks swim, etc.). Moving across representational spaces is possible when relations are abstracted unifying observable differences, such that flying, walking, and swimming is movement in space. Thus, initial premises define the constraints of the mental space in which inference occurs (e.g., birds fly) and premises following specify an application subdomain of this space where property transfer is necessary (e.g., accepting that dogs are birds necessarily implies that dogs fly). Therefore, actual properties (e.g., dogs are not birds) are overwritten by logical constraints connecting mental spaces. Checking consistency of representations in reference to these constraints enables understanding logical fallacies: Accepting “If A then B” does not allow drawing any conclusion about A if only knowing that B occurred (AC) or about B if

only knowing that A did not occur (DA), because B may be caused by causes other than A. That is, the space of the argument is embedded in a context of possibilities, defined by principles integrating inferential rules.

Developing awareness is critical in the formalization of inferential and truth-evaluation rules crystallizing into DLoT. Transition from Bayesian reasoning to rule-based inference at 6–7 years depends on awareness of representations interlinked into predicate–argument structures transforming Bayesian possibilities into emerging logical necessities. Awareness of inferential processes at 8–9 years allows us to represent rules underlying inference and differentiate between them, according to logical operators, engendering biconditionality. Awareness of the relations between rules at 11–13 years allows inducing principles defining relations between rules, involving an awareness of inferential promiscuity enabling conception of an infinite number of alternative premises (Kazi, Kazali, Makris, Spanoudis, & Demetriou, 2019). Hence, PLoT develops in parallel with syllogistic reasoning, with the second formalizing Bayesian principles into logical schemes. Children learn complex concepts by running probabilistic inductions over representations of the world. These inductions and their associated inferential processes are represented with increasing accuracy with development. These representations crystallize inferential imperatives into schemes of syllogistic reasoning. Awareness is a fundamental mechanism in this crystallization raising system 1 into system 2 reasoning. In actual life they often interchange in use akin to a system 1.5 where both Bayesian and logical rules are used as it currently fits.

3. Multiple LoTs

The target article alludes (sect. 6) that there may be more than one LoT. Multiple LoTs are preferable over one LoT. They are established with development, such as a mathematical, a causal, a spatial, and a social LoT, reflecting ability to use multiple symbolic systems obeying different syntaxes. Objects and relations in different domains, such as causal, quantitative, spatial, and social relations, generate different types of symbols and different rules for their transformation. These rules reflect specificities in search, encoding, and evaluation of relations in each domain akin to the differentiation of a common LoT into largely autonomous LoTs. These express the rules and constraints for representing and processing relations specific to each domain (Demetriou et al., 2023). Dehaene, Roumi, Lakretz, Planton, and Sablé-Meyer (2022) concur, proposing multiple LoTs, “akin to computer languages, which encode and compress structures in various domains (mathematics, music, shape...)” (p. 1). In development, these languages diverge in the fashion that Indo-European languages emerged from a common protolanguage, often becoming mutually unintelligible, such as an LoT of mathematics, music, chemistry, and so on. Translating them into each other is possible drawing on a common protoLoT, but it requires special learning. Assuming different LoTs accounts for intra- and interindividual differences in cognitive development and learning difficulties, such as dyslexia or dyscalculia. Therefore, specifying a developing general LoT and local LoTs helps integrate different disciplines, such as cognitive, developmental, and brain science and enables mapping human development on artificial intelligence (Demetriou et al., 2021).

Acknowledgment. Thanks are extended to George Spanoudis for his comments on an earlier version of this article.

Competing interest. None.

References

- Christoforides, M., Spanoudis, G., & Demetriou, A. (2016). Coping with logical fallacies: A developmental training program for learning to reason. *Child Development, 87*, 1856–1876. <https://doi.org/10.1111/cdev.12557>
- Dehaene, S., Roumi, F. A., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: A hypothesis about human singularity. *Trends in Cognitive Science, 26*(9), 751–766. <https://doi.org/10.1016/j.tics.2022.06.010>
- Demetriou, A., Golino, H., Spanoudis, G., Maris, N., & Greiff, S. (2021). The future of intelligence: The central meaning-making unit of intelligence in the mind, the brain, and artificial intelligence. *Intelligence, 87*, 101562. <https://doi.org/10.1016/j.intell.2021.101562>
- Demetriou, A., Makris, N., Kazi, S., Spanoudis, G., & Shayer, M. (2018). The developmental trinity of mind: Cognizance, executive control, and reasoning. *WIREs Cognitive Science, 9*(4), e1461. <https://doi.org/10.1002/wcs.1461>
- Demetriou, A., Spanoudis, G., Christou, C., Greiff, S., Makris, N., Vainikainen, M. P., & Gonida, E. (2023). Cognitive and personality predictors of school performance from preschool to secondary school: An overarching model. *Psychological Review, 130*, 480–512. <https://doi.org/10.1037/rev0000399>
- Kazi, S., Kazali, E., Makris, N., Spanoudis, G., & Demetriou, A. (2019). Cognizance in cognitive development: A longitudinal study. *Cognitive Development, 52*, 100805. <https://doi.org/10.1016/j.cogdev.2019.100805>
- Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology, 71*, 305–330. <https://doi.org/10.1146/annurev-psych-010419-051132>

Linguistic structure and the languages-of-thought

Gabe Dupre 

School of Political, Global, and Social Studies, Keele University,
Staffordshire, UK
g.g.dupre@keele.ac.uk
<https://gabedupre.weebly.com/>

doi:10.1017/S0140525X23002054, e274

Abstract

Quilty-Dunn et al. adopt a methodology for psychology connecting behavioral capacities to the format of the mental systems underlying them. This methodology opens up avenues connecting linguistic theory to comparative psychology. On the assumption that language structures thought, identifying the formal structure of human language can generate hypotheses connecting distinctively human cognitive traits to the distinctive structures of human language.

Quilty-Dunn et al. identify a cluster of traits (discrete constituents, predicate–argument structure, role-filler independence, logical operators, inferential promiscuity, and abstract conceptual content) characteristic of “language-like” psychological states and processes, and marshal a wide range of empirical evidence that seems best explained by the positing of psychological systems with these properties. As they note, however, there are many different ways that mental systems could exemplify these properties (see also Mandelbaum et al., 2022). Specifically, even within the genus of “language-like” systems, there are a wide variety of possible specific formal structures, or formats. The methodology they endorse for identifying and classifying mental systems involves identifying an organism’s behavioral and cognitive capacities, and seeing which sort of mental format would best account for these. Although they are rightly keen to stress the difference

between claims that some organism thinks in a language-of-thought and that this organism thinks in natural (i.e., human) language, and of course to avoid quibbling about whether some “language-like” system is *really* a language, this opens up the possibility of explaining a range of specifically human cognitive capacities by appeal to the apparently unique formats made available by natural language. In this brief comment I will point in some suggestive directions along these lines.

The cluster of traits identified by Quilty-Dunn et al. seems most apt to characterize systems with roughly the structure of predicate logic. They specify an n -place predicate, and n arguments, generating the traditional philosopher’s notion of a proposition, which can then serve as an input for further combination and manipulation, such as logical inference. From the perspective of linguistic theory, such structures are more closely analogous to a verb phrase (VP), the domain of lexical content, rather than a complete sentential clause. A fairly widespread, although controversial, view in generative linguistics (see, e.g., Wiltschko, 2014) is that in addition to the lexical domains which specify, roughly, events and their participants, human linguistic structures contain a “functional spine,” the locus of a range of linguistic features including inflection, mood, force, and more. If these aspects of linguistic structure are indeed distinctive of human language, this raises the possibility that we might be able to appeal to them in explaining aspects of human cognition not found elsewhere in the animal kingdom, along the explanatory lines described by Quilty-Dunn et al. Where we find distinctive formal structure, we can seek distinctive cognitive capacities to be explained.

Of course, linguists posit such structures precisely to appeal to distinctive human cognitive and behavioral capacities involving our use of language. But, if certain hypotheses connecting human language to human thought more generally are along the right lines (e.g., Carruthers, 2002; Chomsky, Gallego, & Ott, 2019; Dupre, 2020), our explanatory reach may be greater, and we may be able to explain distinctively human, but intuitively nonlinguistic, capacities by appeal to the mental structures made available by our linguistic faculty.

Consider, for example, the inflectional phrase (IP), one of the most prominent constituents of the functional spine. On one standard view, the primary function of IP is to “anchor” the event description provided by the VP to features of the discourse (see, e.g., Enç, 1987; Ritter & Wiltschko, 2014). Most commonly, this involves tense-marking, locating the described events in time, relative to the time of conversation, but other options appear to be available, anchoring described events spatially or relative to conversational participants (Ritter & Wiltschko, 2009). Such anchoring appears to be required by the structures and operations made available by the language faculty, even in superficially tenseless languages (see, e.g., Matthewson, 2006; Sybesma, 2007).

If, and these are big “ifs,” human thought is structured by human language, and if human language requires anchoring, this is suggestive of a unification of language and one of the other allegedly unique capacities of the human mind, namely “mental time travel.” No other animal has uncontroversially demonstrated the ability to associate specific event-type representations with times and individuals the way humans do in episodic memory (see, e.g., Roberts & Feeney, 2009; Hoerl & McCormack, 2017, for reviews). If the structures of nonhuman cognition are well-characterized by the propositional structures described by Quilty-Dunn et al., although human thoughts are structured by the functional hierarchy posited by generative grammarians, this could go some way to turning two unique features of

human cognition into one: Our ability to anchor our memories to specific temporal windows may be, or be causally/developmentally related to, our ability to form linguistic structures with an IP serving precisely this function.

Of course, much more work would need to be done to turn this suggestive similarity into a substantiated empirical hypothesis. And all of the work I have appealed to here is highly controversial. But I believe that the framework for psychological explanation provided by Quilty-Dunn et al. provides a highly productive way to bring to bear the results of contemporary linguistic theory onto questions in comparative psychology, in this case and a wide range of others.

Financial support. Funding for this research was provided by the Leverhulme Trust (Ref: ECF-2020-424).

Competing interest. None.

References

- Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, 25(6), 657–674.
- Chomsky, N., Gallego, A. J., & Ott, D. (2019). Generative grammar and the faculty of language: Insights, questions, and challenges. *Generative Syntax. Questions, Crossroads, and Challenges [special issue:] Catalan Journal of Linguistics*, 229–261.
- Dupre, G. (2020). What would it mean for natural language to be the language of thought? *Linguistics and Philosophy*, 44, 773–812.
- Enç, M. (1987). Anchoring conditions for tense. *Linguistic Inquiry*, 18, 633–657.
- Hoerl, C., & McCormack, T. (2017). Animal minds in time: The question of episodic memory. In K. Andrews & J. Beck (Eds.) *The Routledge handbook of philosophy of animal minds* (pp. 56–64). Routledge.
- Mandelbaum, E., Dunham, Y., Feiman, R., Firestone, C., Green, E. J., Harris, D., ... Quilty-Dunn, J. (2022). Problems and mysteries of the many languages of thought. *Cognitive Science*, 46(12), e13225.
- Matthewson, L. (2006). Temporal semantics in a superficially tenseless language. *Linguistics and Philosophy*, 29(6), 673–714.
- Ritter, E., & Wiltschko, M. (2009). Varieties of INFL: Tense, location, and person. *Alternatives to Cartography*, 153, 202.
- Ritter, E., & Wiltschko, M. (2014). The composition of INFL: An exploration of tense, tenseless languages, and tenseless constructions. *Natural Language & Linguistic Theory*, 32, 1331–1386.
- Roberts, W. A., & Feeney, M. C. (2009). The comparative study of mental time travel. *Trends in Cognitive Sciences*, 13(6), 271–277.
- Sybesma, R. (2007). Whether we tense-agree overtly or not. *Linguistic Inquiry*, 38(3), 580–587.
- Wiltschko, M. (2014). *The universal structure of categories* (Vol. 142). Cambridge University Press.

On the hazards of relating representations and inductive biases

Thomas L. Griffiths^a , Sreejan Kumar^b
and R. Thomas McCoy^c

^aDepartments of Psychology and Computer Science, Princeton University, Princeton, NJ, USA; ^bNeuroscience Institute, Princeton University, Princeton, NJ, USA and ^cDepartment of Computer Science, Princeton University, Princeton, NJ, USA
tomg@princeton.edu
sreejank@princeton.edu
tom.mccoy@princeton.edu
<http://cocosci.princeton.edu/tom/>
<http://sreejankumar.com>
<https://rtmccoy.com/>

doi:10.1017/S0140525X23002042, e275

Abstract

The success of models of human behavior based on Bayesian inference over logical formulas or programs is taken as evidence that people employ a “language-of-thought” that has similarly discrete and compositional structure. We argue that this conclusion problematically crosses levels of analysis, identifying representations at the algorithmic level based on inductive biases at the computational level.

Over the last few decades probabilistic models of cognition, which explain human behavior in terms of Bayesian inference over a set of hypotheses, have been applied to a wide range of phenomena (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010). But what does the success of a particular probabilistic model in capturing human behavior imply? Probabilistic models of cognition are typically defined at Marr’s (1982) “computational” level, characterizing the abstract problems human minds have to solve and their ideal solutions. More precisely, they characterize the ideal solutions to inductive problems, where an agent has to draw conclusions that go beyond the available data. The content of a probabilistic model of cognition, expressed via the set of hypotheses and their prior probabilities, is a claim about the *inductive biases* that guide such inferences – those factors other than the data that influence the hypothesis the agent selects (Mitchell, 1997). Here, we argue that drawing conclusions that go beyond these inductive biases – and in particular, inferring support for specific cognitive processes and representations – can be problematic.

Inductive biases are at a different level of analysis from cognitive processes and representations, which Marr (1982) located at the “representation and algorithm” level. Representations and algorithms are notoriously underdetermined by observable data (Anderson, 1978). This underdetermination motivated Anderson (1990) to develop rational analysis, the approach adopted in almost all applications of probabilistic models of cognition. This approach explicitly focuses on abstract problems and their ideal solutions rather than the processes and representations that implement them (inspiring critiques, e.g., Jones & Love, 2011). Probabilistic models need some way of representing hypotheses, but such representations do not necessarily guide human behavior. Rather, they are theoretical constructs that help scientists describe inductive biases.

Finding that a particular inductive bias seems to characterize human behavior places constraints on the representations and algorithms that might be involved, but those constraints rarely pick out a unique solution. To give a simple example, consider the problem of learning a linear relationship between two variables. A probabilistic model identifies a set of hypotheses (e.g., all linear functions), defines a prior distribution over those hypotheses, and then performs Bayesian inference. This kind of solution could be implemented by an agent that explicitly represents a set of linear functions and uses an algorithm to update its beliefs about the posterior probability of each hypothesis as new data are observed (see, e.g., Sanborn, Griffiths, & Navarro, 2010). The behavior of this agent will match that of the ideal Bayesian model. However, an agent that seems quite different – a neural network with one hidden layer and a linear output function that updates its weights by a few iterations of gradient descent – will also produce an answer that matches that Bayesian model (assuming a Gaussian prior; see Santos, 1996). Two very different representations and algorithms are consistent with the same

computational-level account (for a real modeling example, see Feldman, Griffiths, & Morgan, 2009).

Quilty-Dunn et al. argue from the success of probabilistic models of cognition based on Bayesian inference over logical formulas and programs to the conclusion that people employ a similarly discrete and compositional “language-of-thought.” This argument crosses levels of analysis in the same problematic way. What we are licensed to conclude from the success of these models is that logical formulas and programs are useful in characterizing human inductive biases for certain problems, not that humans use these representations when solving those problems. Any stronger conclusion seems particularly problematic in light of the recent successes of deep neural networks that Quilty-Dunn et al. mention, because these systems may not require discrete or compositional representations. Metalearning – training a system to perform a set of related tasks – provides a way to create neural networks with specific inductive biases, and has formal connections to learning a prior for Bayesian inference (Grant, Finn, Levine, Darrell, & Griffiths, 2018). Metalearning has been used to train neural networks to perform tasks characterized at the computational level by Bayesian models based on symbolic representations, such as theory-of-mind (Rabinowitz et al., 2018) and causal learning (Dasgupta et al., 2019). Analysis of the internal representations of related systems shows that they contain information that can be used to reconstruct appropriate posterior distributions (Mikulik et al., 2020). It thus seems plausible that such systems might produce behavior that is just as consistent with Bayesian inference over logical formulas and programs as that of humans.

The possible existence of deep neural networks that can be analyzed at the computational level in terms of Bayesian inference blocks strong conclusions about the language-of-thought, as the representations learned by these networks could emulate the associated behavior without requiring discreteness. Our investigations of networks trained by metalearning show that they can emulate human performance on abstract tasks without explicit representations of the relevant abstractions (Kumar et al., 2022). In some cases, deep neural networks succeed on abstract tasks by learning compositionally structured representations (McCoy, Linzen, Dunbar, & Smolensky, 2019), but these representations remain continuous, making them importantly different from the inherently discrete ones postulated in the language-of-thought hypothesis (Smolensky, 1988). Such results align with theoretical work showing that compositional behavior does not require discrete representations (Smolensky, McCoy, Fernandez, Goldrick, & Gao, 2022). Indeed, the best current models of language itself – which is the prototypical example of a compositional domain (Pinker & Prince, 1988), as suggested by its use in the name *language-of-thought* – are deep networks that have continuous internal representations (e.g., Chowdhery et al., 2022).

Algorithms and representations may not be identifiable, but we can at least narrow down the equivalence class of possibilities through careful experimentation – behavioral work focused on response times and errors, neuroscientific studies of what the brain might be encoding, and computational simulations – designed to provide strong tests of alternative hypotheses. Until we can definitively do so, the fact that a discrete, compositional language-of-thought is useful as an abstract way of characterizing human inductive biases still allows the possibility that the actual representations and algorithms underlying human cognition may have a very different character.

Acknowledgment. We thank Tania Lombrozo for helpful comments.

Financial support. This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship under Grant No. 2204152 and the Office of Naval Research under Grant No. N00014-18-1-2873. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Office of Naval Research.

Competing interest. None.

References

- Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, 85(4), 249.
- Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... Fiedel, N. (2022). PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., ... Kurth-Nelson, Z. (2019). Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*.
- Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4), 752.
- Grant, E., Finn, C., Levine, S., Darrell, T., & Griffiths, T. (2018). Recasting gradient-based meta-learning as hierarchical Bayes. In *International Conference on Learning Representations*. OpenReview.
- 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.
- Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34(4), 169–188.
- Kumar, S., Correa, C. G., Dasgupta, I., Marjeh, R., Hu, M., Hawkins, R. D. (2022). *Using natural language and program abstractions to instill human inductive biases in machines*. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (pp. 167–180). Curran Associates.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman.
- McCoy, R. T., Linzen, T., Dunbar, E., & Smolensky, P. (2019). RNNs implicitly implement tensor product representations. In *International Conference on Learning Representations*. OpenReview.
- Mikulik, V., Delétang, G., McGrath, T., Genewein, T., Martic, M., Legg, S., & Ortega, P. (2020). Meta-trained agents implement Bayes-optimal agents. *Advances in Neural Information Processing Systems*, 33, 18691–18703.
- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.
- Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1–2), 73–193.
- Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. In J. Dy & A. Krause (Eds.), *International Conference on Machine Learning* (pp. 4218–4227). PMLR.
- Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4), 1144.
- Santos, R. J. (1996). Equivalence of regularization and truncated iteration for general ill-posed problems. *Linear Algebra and its Applications*, 236, 25–33.
- Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1), 1–23.
- Smolensky, P., McCoy, R., Fernandez, R., Goldrick, M., & Gao, J. (2022). Neurocompositional computing: From the central paradox of cognition to a new generation of AI systems. *AI Magazine*, 43(3), 308–322.

Advanced testing of the LoT hypothesis by social reasoning

David J. Grüning^{a,b} 

^aPsychology Department, Heidelberg University, Heidelberg, Germany and
^bDepartment of Survey Design and Methodology, GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany
david.gruening@psychologie.uni-heidelberg.de

doi:10.1017/S0140525X2300184X, e276

Abstract

I elaborate on Quilty-Dunn et al.'s integration of the language-of-thought hypothesis in social reasoning by outlining two discrepancies between the experimental paradigms referred to by the authors and the social world: Self-referential projection and deliberate thinking in experiments. Robust tests of the hypothesis in social reasoning should include observational, natural, and cross-cultural approaches.

I aim to elaborate on Quilty-Dunn et al.'s illustrative argumentation in support of the language-of-thought (LoT) hypothesis through the social psychological lens presented in the last section of their article. Although the authors' address of conflict problems (target article, sect. 6.1) and implicit attitude research (target article, sect. 6.2) is compelling in its experimental context, it is not far-reaching enough, that is, it is only weakly informative about the actual socially embedded reasoning of individuals. What is missing is a conceptual test of the LoT hypothesis in real-life social situations of logical reasoning – for example, such situations prevalent in research on competitive or collaborative games and strategic thinking (e.g., Colman, 2003; Grüning & Krueger, 2021, 2022; Hedden & Zhang, 2002). Realistic social situations of reasoning are different to the cases addressed by Quilty-Dunn et al. in several aspects. In this commentary, I outline two aspects in more detail: (1) social and self-referential projection, and (2) deliberate thinking through experimental artificiality.

First, social situations of logical reasoning are highly complicated by experiential social learning and self-referential projection. For illustration let us turn to an example: Quilty-Dunn et al. iterate an experiment by Kurdi and Dunham (2021) in which participants were presented with, among other statements, the following simple logical statement: “If you see a green circle, you can conclude that Ibbonif is malicious” (target article, sect. 6.2, para. 3). Adapting this straightforward statement to a context of social inference – for instance: “If you see a smirk on the face, you can conclude that Peter is malicious,” – can quickly ascend individuals into a rabbit hole of applying their (1) own social learning and (2) induction from introspection about their self. Both confounds substantially with learning and testing phases as presented in the original study. For one, the absence of a smile is not as unambiguously informative as the absence of the green circle in Kurdi and Dunham's (2021) experiment. Social cues, like facial expressions, are predominantly multicausal and, hence, ambiguous in their information about the real state of the world, more so the less contextual information is provided (e.g., “Peter is smiling after something has happened.” vs. “[...] after something terrible has happened.”). An individual can never “conclude” with full certainty what a social signal informs one about. Second, when evaluating real social situations, individuals cannot step away from using themselves as referential source of information to evaluate the situation. In the asocial situation of coloured circles and Ibbonifs, self-referential inference is not applicable, unless in the unlikely event that an individual draws connections between these concepts and their self and personal experiences. However, in social situations, that is, situations including other people in interaction, self-referentiality is a very prominent strategy for social reasoning (e.g., Krueger, 2008, 2013; Krueger & Grüning, 2021; Krueger, Grüning, & Heck, 2023). Both of the here outlined complications occur when we move from quasi-

social to in-fact social statements. They are intended to illustrate that simple cases where associative (i.e., social learning) and propositional logic are easily distinguishable, and self-referential projection is no confound are difficult to find in actual social reasoning.

Second, the high artificiality of the experimental context and task in both of the authors' research examples should be taken into account when interpreting their results as evidence for a specific reasoning hypothesis. The experimental context itself is expected to increase participants' cognitive alertness and motivation for accuracy (e.g., Orne, 1962; Zizzo, 2010). The artificiality of most experimental reasoning tests, including the authors' examples, is further likely to encourage participants' deliberate instead of intuitive thinking regarding reasoning statements (see, as process explanation; Evans, 2008; Kahneman, 2011; recently, De Neys, 2022) as stimulus materials. In this respect, a strong interpretation of the discussed experiments might commit the same fallacy as early interpretations of human bias (e.g., Kahneman, Slovic, & Tversky, 1982) that were later challenged to contain experimental artefacts (e.g., Gigerenzer, 1996; Hertwig, Leucker, Pachur, Spiliopoulos, & Pleskac, 2022; but also see, Vranas, 2000). I hasten to note that this is largely an inherent problem of the experimental context created by conversational norms and the idiosyncrasy of the experimental design (e.g., Schwarz, 1994, 1999), not a shortcoming by the authors. Experimental exploration is, by all means, meaningful. However, at the same time, it is just a first step to investigate a psychological phenomenon, even more so when considering social cognition phenomena like social reasoning. The experimental artificiality can be fled by also using observational and field study designs, exchanging some internal for ecological and external validity. Before the experiments, that Quilty-Dunn et al. call upon to argue for LoT in the social psychological space, have been extended to more ecologically valid contexts, generalizable claims of any sort, including the LoT hypothesis, should be modest.

Concluding, I welcome Quilty-Dunn et al.'s attempt for an exhaustive integration of the LoT hypothesis in psychological theory and empirics. Relevantly, with my commentary I do not attempt to rebut or support the LoT hypothesis. I seek to make the authors and readers aware of the fact that for a robust, that is, a persuasive, test of the LoT hypothesis in the social context, researchers cannot exclusively revert to simple experimental imitations of social reasoning. Instead, existing findings from realistic social inference-making scenarios have to be considered by the authors and observational and field experimental approaches need to be focused on in the future. Cross-cultural exploration, as an advanced extension of social psychology, would provide an additional opportunity to test the generalizability of the LoT hypothesis.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

- Colman, A. M. (2003). Depth of strategic reasoning in games. *Trends in Cognitive Sciences*, 7(1), 2–4. [https://doi.org/10.1016/S1364-6613\(02\)00006-2](https://doi.org/10.1016/S1364-6613(02)00006-2)
- De Neys, W. (2022). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 1–68. <https://doi.org/10.1017/S0140525X2200142X>
- Evans, J. St. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255–278. <https://doi.org/10.1146/annurev.psych.59.103006.093629>

- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, 103(3), 592–596. <https://doi.org/10.1037/0033-295X.103.3.592>
- Grüning, D. J., & Krueger, J. I. (2021). Strategic thinking: A random walk into the rabbit hole. *Collabra: Psychology*, 7(1), 24921. <https://doi.org/10.1525/collabra.24921>
- Grüning, D. J., & Krueger, J. I. (2022). Strategic thinking in the shadow of self-enhancement: Benefits and costs. *PsyArXiv*. <https://doi.org/10.31234/osf.io/gtc2m>
- Hedden, T., & Zhang, J. (2002). What do you think I think you think?: Strategic reasoning in matrix games. *Cognition*, 85(1), 1–36. [https://doi.org/10.1016/S0010-0277\(02\)00054-9](https://doi.org/10.1016/S0010-0277(02)00054-9)
- Hertwig, R., Leucker, C., Pachur, T., Spiliopoulos, L., & Pleskac, T. J. (2022). Studies in ecological rationality. *Topics in Cognitive Science*, 14(3), 467–491. <https://doi.org/10.1111/tops.12567>
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.
- Krueger, J. I. (2008). From social projection to social behaviour. *European Review of Social Psychology*, 18(1), 1–35. <https://doi.org/10.1080/10463280701284645>
- Krueger, J. I. (2013). Social projection as a source of cooperation. *Current Directions in Psychological Science*, 22(4), 289–294. <https://doi.org/10.1177/0963721413481352>
- Krueger, J. I., & Grüning, D. J. (2021). Psychological perversities and populism. In J. P. Forgas, W. D. Crano, & K. Fiedler (Eds.), *The psychology of populism* (pp. 125–142). Routledge.
- Krueger, J. I., Grüning, D. J., & Heck, P. R. (2023). Inductive reasoning model. *PsyArXiv*. <https://doi.org/10.31234/osf.io/3yasf>
- Kurdi, B., & Dunham, Y. (2021). Sensitivity of implicit evaluations to accurate and erroneous propositional inferences. *Cognition*, 214, 104792. <https://doi.org/10.1016/j.cognition.2021.104792>
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11), 776–783. <https://doi.org/10.1037/h0043424>
- Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 26, pp. 123–162). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60153-7](https://doi.org/10.1016/S0065-2601(08)60153-7)
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54(2), 93–105. <https://doi.org/10.1037/0003-066X.54.2.93>
- Vranas, P. B. (2000). Gigerenzer's normative critique of Kahneman and Tversky. *Cognition*, 76(3), 179–193. [https://doi.org/10.1016/S0010-0277\(99\)00084-0](https://doi.org/10.1016/S0010-0277(99)00084-0)
- Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13, 75–98. <https://doi.org/10.1007/s10683-009-9230-z>

Compositionality in visual perception

Alon Hafri^a , E. J. Green^b and Chaz Firestone^c 

^aDepartment of Linguistics and Cognitive Science, University of Delaware, Newark, DE, USA; ^bDepartment of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, MA, USA and ^cDepartment of Psychological and Brain Sciences, Johns Hopkins University and Baltimore, MD, USA
alon@udel.edu; <https://pal.lingcogsci.udel.edu/>
ejgr@mit.edu; <https://sites.google.com/site/greenedwinj/>
chaz@jhu.edu; <https://perception.jhu.edu/>

doi:10.1017/S0140525X23001838, e277

Abstract

Quilty-Dunn et al.'s wide-ranging defense of the Language of Thought Hypothesis (LoTH) argues that vision traffics in abstract, structured representational formats. We agree: Vision, like language, is *compositional* – just as words compose into phrases, many visual representations contain discrete constituents that combine in systematic ways. Here, we amass evidence extending this proposal, and explore its implications for how vision interfaces with the rest of the mind.

The world we see is populated by colors, textures, edges, and countless other visual features. Yet we see more than a

collection of features: We also see whole objects, and relations within and between those objects. How are these entities represented? Here, we advance the case for LoT-like representation in perception. We argue that at least two types of visual representations are compositional, and we explore their connections with the rest of the mind.

Consider the hands in Figure 1A. Although they differ in various superficial features, they appear to share something: their *structure* – specifically, their *skeletal structure*. The same parts are connected in the same ways, just in different poses. Similarly, the middle shape in Figure 1B shares its structure with the left shape but not the right shape, even though the middle and right shapes share other features. Skeletal representations describe shapes via their parts’ intrinsic axes and connections, often in a hierarchical *tree* format, wherein certain parts “descend” or “offshoot” from others (Feldman & Singh, 2006). Copious evidence suggests that skeletal representations are psychologically real, implicated in detection (Kovács & Julesz, 1994; Wilder, Feldman, & Singh, 2016), discrimination (Lowet, Firestone, & Scholl, 2018), categorization (Wilder, Feldman, & Singh, 2011), aesthetics (Van Tonder, Lyons, & Ejima, 2002), and more (Firestone & Scholl, 2014; Psotka, 1978).

We contend that skeletal representations exhibit several of Quilty-Dunn et al.’s LoT properties: *Discrete constituents*, *role-filler independence*, and *abstract content*. First, skeletal representations contain discrete constituents that represent axis structure independently of surrounding boundaries, composing with boundary representations to describe overall shape. This may explain why infants (Ayzenberg & Lourenco, 2022) and adults (Wilder et al., 2011) categorize novel shapes by skeletal structure despite differences in surface properties. Second, representations of individual parts exhibit role-filler independence, retaining identity over changes in position within the overall skeletal representation. Such *transportability* (Fodor, 1987) explains why we can easily determine when distinct shapes share the same parts, and why

such shapes prime one another (Cacciamani, Ayars, & Peterson, 2014). Third, skeletal representations are abstract, expressing aspects of shape that appear stable despite part articulations (Fig. 1A), changes in surface properties (Fig. 1B; Green, 2019), and sense modality (Green, 2022). Moreover, visual brain areas encode skeletal structure across surface changes (Ayzenberg, Kamps, Dilks, & Lourenco, 2022; Hung, Carlson, & Connor, 2012; Lescroart & Biederman, 2013). Skeletal representations may also encode nonmetric, categorical properties – for example, *straight/curved* and *symmetric/asymmetric* (Amir, Biederman, & Hayworth, 2012; Green, 2017; Hafri, Gleitman, Landau, & Trueswell, 2023).

We suggest that these LoT properties make skeletal representations compositional: Discrete constituents encoding different geometrical elements and properties combine to form representations of global shape.

Compositionality in vision extends to relations *between* objects. Consider the object pairs in Figure 1C. They appear to share something: the relation *containment*. Visual processing respects this commonality – it represents relations between objects, beyond the objects themselves (Hafri & Firestone, 2021). Such representations also exhibit several LoT properties. First, visual processing represents relations abstractly and categorically: Observers are more sensitive to metric changes across relational category boundaries (e.g., from containing to merely touching) than within (e.g., from one instance of containment to another; Lovett & Franconeri, 2017), and even “confuse” instances of the same relation for one another (Hafri, Bonner, Landau, & Firestone, 2020). Furthermore, visual brain areas encode eventive relations abstractly, generalizing across event participants (Hafri, Trueswell, & Epstein, 2017; Wurm & Lingnau, 2015).

Second, such representations contain discrete constituents and exhibit role-filler independence, in ways that augment Quilty-Dunn et al.’s discussion. Consider Figure 1D. Both images

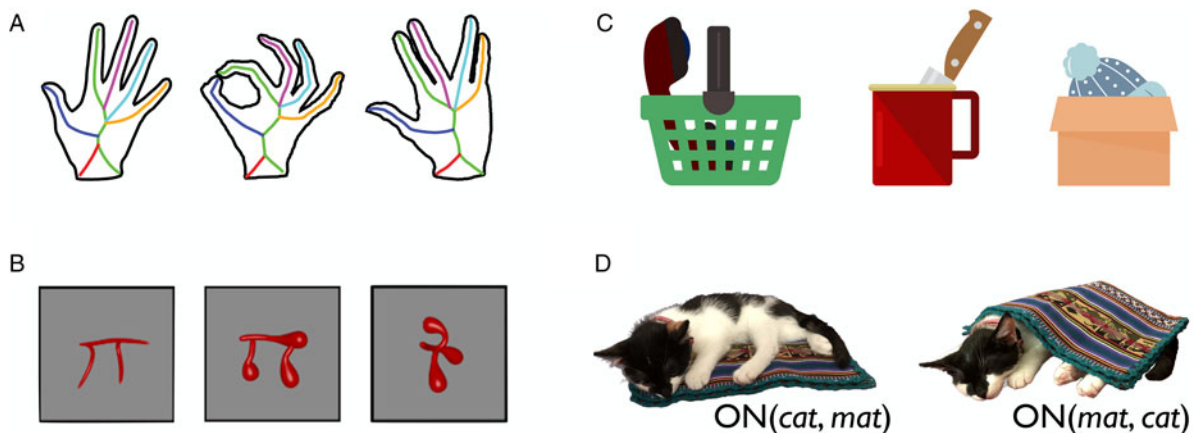


Figure 1 (Hafri et al.). Demonstrations of compositionality in visual perception. (A) The three hands shown here differ in global shape, the locations of their boundaries, and other surface features; however, they appear to share something: Their *structure* – specifically, their *skeletal structure* (indicated by the inset colored lines). The same parts have taken on different poses. Skeletal shape representations describe objects in terms of the axes of their parts, including how those parts are arranged with respect to one another, in ways that instantiate several core LoT properties. (Adapted from Lowet et al., 2018.) (B) Skeletal shape representations explain why infants and adults can see that the middle shape shares something with the leftmost shape that it does not share with the rightmost shape, even though the middle and rightmost shapes share other features. (Adapted from Ayzenberg & Lourenco, 2019.) (C) The three object pairs shown here differ in a variety of visual features, and even involve different objects – but each seems to instantiate the same relation: *containment*. Recent evidence suggests that the mind rapidly and automatically encodes such relations, representing the relation itself separately from the objects participating in it. (Adapted from Hafri et al., 2020.) (D) These two images depict the same objects (cat and mat) and the same relation (*support*), but differ in their *structure* – a cat on a mat is a very different scene from a mat on a cat. Put differently, “argument order” matters: $R(x,y)$ may be quite different than $R(y,x)$, and there is evidence that visual processing is sensitive to this difference in compositional structure. (Adapted from Hafri & Firestone, 2021.)

involve the same objects (cat and mat) and relation (*support*), but cat-on-mat differs from mat-on-cat in compositional structure. Thus, “argument order” matters – the “fillers” map to different roles. Recent work shows that vision is sensitive to this difference. When observers repeatedly reported the location of a target individual (e.g., blue-shirted man) in a stream of action photographs (e.g., blue-kicking-red, red-pushing-blue), a “switching cost” emerged: Slower responses when the target individual’s role (*Agent/Patient*) switched (e.g., pusher on trial $n - 1$ but kickee on trial n), suggesting that observers encoded relational structure automatically (Hafri, Trueswell, & Strickland, 2018).

These properties make representations of categorical between-object relations compositional: Discrete constituents encoding entities and relations combine to form representations of structured situations.

The prospect of LoT-like, compositional visual representations impacts broader debates about perception’s *format*. Many claim that perceptual representations are constitutively iconic, analog, or “picture-like” (Burge, 2022; Carey, 2009; Dretske, 1981; Kosslyn, Thompson, & Ganis, 2006). However, although LoT-like formats clearly suffice to encode categorical, nondegreed relations (e.g., containment), many iconic formats may not – particularly accounts requiring perceptual icons to mirror graded degrees of difference in perceptible properties (e.g., orientation or brightness; Block, 2023).

This perspective also raises exciting questions and research directions. For example, it may partially explain how information from perception is “readily consumed” by cognitive and linguistic systems (because of the similar formats of some perceptual and higher-level representations; Cavanagh, 2021; Quilty-Dunn, 2020). Recent work explores these connections explicitly: Skeletal shape representations impact aesthetic preferences and linguistic descriptions of shapes (Sun & Firestone, 2022a, 2022b), and representations of symmetry and roles may be shared across perception and language (Hafri et al., 2018, 2023; Rissman & Majid, 2019; Strickland, 2017). One could also investigate the “psychophysics” of compositional processes – the timing and ordering of how relational representations are built from their parts.

Nevertheless, LoT-like perceptual representations may not be *fully* language-like. Although perception plausibly predicates properties of individuals (Quilty-Dunn & Green, 2023), it may lack the full *expressive freedom* of first-order logic (Camp, 2018), especially logical connectives needed for truth-functional completeness (Mandelbaum et al., 2022). Perception may be able to represent that an object is red but not that it is *not* red. Moreover, certain perceptual formats may impose constraints on which properties are attributable to which individuals – constraints absent from higher-level cognition. Perhaps perception cannot explicitly represent relations between nonadjacent object parts, or eventive relations of long durations (e.g., a jack slowly lifting a car).

Because perception and thought confront multifarious tasks with different computational demands, we contend that they comprise a multiplicity of formats (Marr, 1982; Yousif, 2022), each optimized for different computations, and some more LoT-like than others. Thus, any theory positing a single-privileged format for perception or thought should be met with suspicion. Instead, researchers should heed Quilty-Dunn et al.’s advice to “let a thousand representational formats bloom” (target article, sect. 2, para. 2).

Acknowledgments. For comments on an earlier draft, the authors acknowledge members of the JHU Perception and Mind Laboratory.

Financial support. This study was supported by NSF BCS no. 2021053 awarded to C. F.


Competing interest. None.

References

- Amir, O., Biederman, I., & Hayworth, K. J. (2012). Sensitivity to nonaccidental properties across various shape dimensions. *Vision Research*, 62, 35–43.
- Ayzenberg, V., Kamps, F. S., Dilks, D. D., & Lourenco, S. F. (2022). Skeletal representations of shape in the human visual cortex. *Neuropsychologia*, 164, 108092.
- Ayzenberg, V., & Lourenco, S. F. (2019). Skeletal descriptions of shape provide unique perceptual information for object recognition. *Scientific Reports*, 9, 1–13.
- Ayzenberg, V., & Lourenco, S. F. (2022). Perception of an object’s global shape is best described by a model of skeletal structure in human infants. *eLife*, 11, e74943.
- Block, N. (2023). *The border between seeing and thinking*. Oxford University Press.
- Burge, T. (2022). *Perception: First form of mind*. Oxford University Press.
- Cacciamani, L., Ayars, A. A., & Peterson, M. A. (2014). Spatially rearranged object parts can facilitate perception of intact whole objects. *Frontiers in Psychology*, 5, 482.
- Camp, E. (2018). Why maps are not propositional. In A. Grzankowski & M. Montague (Eds.), *Non-propositional intentionality* (pp. 19–45). Oxford University Press.
- Carey, S. (2009). *The origin of concepts*. Oxford University Press.
- Cavanagh, P. (2021). The language of vision. *Perception*, 50, 195–215.
- Dretske, F. I. (1981). *Knowledge and the flow of information*. MIT Press.
- Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 18014–18019.
- Firestone, C., & Scholl, B. J. (2014). “Please tap the shape, anywhere you like”: Shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological Science*, 25, 377–386.
- Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press.
- Green, E. J. (2017). A layered view of shape perception. *The British Journal for the Philosophy of Science*, 68, 355–387.
- Green, E. J. (2019). On the perception of structure. *Noûs*, 53, 564–592.
- Green, E. J. (2022). The puzzle of cross-modal shape experience. *Noûs*, 56, 867–896.
- Hafri, A., Bonner, M. F., Landau, B., & Firestone, C. (2020). A phone in a basket looks like a knife in a cup: Role-filler independence in visual processing. *PsyArXiv*. <https://psyarxiv.com/jx4yg>
- Hafri, A., & Firestone, C. (2021). The perception of relations. *Trends in Cognitive Sciences*, 25, 475–492.
- Hafri, A., Gleitman, L. R., Landau, B., & Trueswell, J. C. (2023). Where word and world meet: Language and vision share an abstract representation of symmetry. *Journal of Experimental Psychology: General*, 152, 509–527.
- Hafri, A., Trueswell, J. C., & Epstein, R. A. (2017). Neural representations of observed actions generalize across static and dynamic visual input. *The Journal of Neuroscience*, 37, 3056–3071.
- Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition*, 175, 36–52.
- Hung, C. C., Carlson, E. T., & Connor, C. E. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron*, 74, 1099–1113.
- Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. Oxford University Press.
- Kovács, I., & Julesz, B. (1994). Perceptual sensitivity maps within globally defined visual shapes. *Nature*, 370, 644–646.
- Lescroart, M. D., & Biederman, I. (2013). Cortical representation of medial axis structure. *Cerebral Cortex*, 23, 629–637.
- Lovett, A., & Franconeri, S. L. (2017). Topological relations between objects are categorically coded. *Psychological Science*, 28, 1408–1418.
- Lowet, A. S., Firestone, C., & Scholl, B. J. (2018). Seeing structure: Shape skeletons modulate perceived similarity. *Attention, Perception, & Psychophysics*, 80, 1278–1289.
- Mandelbaum, E., Dunham, Y., Feiman, R., Firestone, C., Green, E. J., Harris, D., ... Quilty-Dunn, J. (2022). Problems and mysteries of the many languages of thought. *Cognitive Science*, 46, e13225.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman.
- Psotka, J. (1978). Perceptual processes that may create stick figures and balance. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 101–111.
- Quilty-Dunn, J. (2020). Concepts and predication from perception to cognition. *Philosophical Issues*, 30, 273–292.
- Quilty-Dunn, J., & Green, E. J. (2023). Perceptual attribution and perceptual reference. *Philosophy and Phenomenological Research*, 106, 273–298. doi: 10.1111/phpr.12847.
- Rissman, L., & Majid, A. (2019). Thematic roles: Core knowledge or linguistic construct? *Psychonomic Bulletin & Review*, 26, 1850–1869.
- Strickland, B. (2017). Language reflects “core” cognition: A new theory about the origin of cross-linguistic regularities. *Cognitive Science*, 41, 70–101.

- Sun, Z., & Firestone, C. (2022a). Beautiful on the inside: Aesthetic preferences and the skeletal complexity of shapes. *Perception*, *51*, 904–918.
- Sun, Z., & Firestone, C. (2022b). Seeing and speaking: How verbal “description length” encodes visual complexity. *Journal of Experimental Psychology: General*, *151*, 82–96.
- Van Tonder, G. J., Lyons, M. J., & Ejima, Y. (2002). Visual structure of a Japanese Zen garden. *Nature*, *419*, 359–360.
- Wilder, J., Feldman, J., & Singh, M. (2011). Superordinate shape classification using natural shape statistics. *Cognition*, *119*, 325–340.
- Wilder, J., Feldman, J., & Singh, M. (2016). The role of shape complexity in the detection of closed contours. *Vision Research*, *126*, 220–231.
- Wurm, M. F., & Lingnau, A. (2015). Decoding actions at different levels of abstraction. *Journal of Neuroscience*, *35*, 7727–7735.
- Yousif, S. R. (2022). Redundancy and reducibility in the formats of spatial representations. *Perspectives on Psychological Science*, *17*, 1778–1793.

Incomplete language-of-thought in infancy

Jean-Rémy Hochmann^{a,b} 

^aCNRS UMR5229 – Institut des Sciences Cognitives Marc Jeannerod, Bron, France and ^bUniversité Lyon 1 Claude Bernard, Lyon, France
hochmann@isc.cnrs.fr
<https://sites.google.com/site/jrhochmann/>

doi:10.1017/S0140525X23001826, e278

Abstract

The view that infants possess a full-fledged propositional language-of-thought (LoT) is appealing, providing a unifying account for infants’ precocious reasoning skills in many domains. However, careful appraisal of empirical evidence suggests that there is still no convincing evidence that infants possess discrete representations of abstract relations, suggesting that infants’ LoT remains incomplete. Parallel arguments hold for perception.

The view that infants possess a propositional language-of-thought (LoT) appeals as a unifying account for precocious physical (Stahl & Feigenson, 2015), logical (Cesana-Arlotti et al., 2018), probabilistic (Denison & Xu, 2010; Téglás, Girotto, Gonzalez, & Bonatti, 2007), and social reasoning (Baillargeon, Scott, & He, 2010; Hamlin, Wynn, & Bloom, 2007; Powell & Spelke, 2013). It suggests continuity along development in the format of human thought. But arguing for such continuity also raises questions. Most, if not all, of the cognitive skills of young infants are also documented in nonhuman species (Engelmann et al., 2022; Krupenye, Kano, Hirata, Call, & Tomasello, 2016), suggesting continuity along evolution. We should thus attribute the same type of thoughts to nonhuman animals and human infants, to animals and human adults. How, then, do we account for animals’ failure to acquire human natural languages and develop unique human cognitive skills? Careful appraisal of the available data and careful experimental designs may instead highlight important discontinuities in the format of thought along both developmental and evolutionary scales, suggesting that a full-fledged LoT, involving all six properties identified by Quilty-Dunn et al. is not yet available to young infants (nor to animals).

I applaud the project of Quilty-Dunn et al. to list specific properties of a propositional LoT and evaluate the presence of these

properties in various subdomains of cognitive science. The strength of the evidence for each property in all domains is however unequal. In particular, before concluding that infants possess a full-fledged LoT, we need to provide evidence for each property, individually, and also investigate the limits of each property. I will focus on the first property, “discrete constituents.” It is the most important, as it is presupposed by most other properties: Roles are attributed to discrete constituents; predication combines discrete constituents; logical operators are conceived as discrete constituents. Contrary to Quilty-Dunn et al., I will argue that, although both perception and infant cognition certainly possess discrete representations of objects and possibly of features, there is no evidence for discrete representations of relations in perception nor in prelexical infants.

Although experimental evidence suggests that perceptual representations of relational events and scenes are generalizable to a certain extent (e.g., Goupil, Papeo, & Hochmann, 2022; Papeo, 2020; see Kominsky & Scholl, 2020, for the limits of those generalizations), there is no evidence that those representations are discrete, dissociated from the object representations. Rather, relations may well be represented by perceptual schema composed of discrete *object* representations. The generalizability can be obtained through the underspecification of object representations, a process we previously called “abstraction by impoverishment” (Hochmann & Papeo, 2021). For instance, in perception, a schematic social interaction would consist in two schematic bodies facing each other (Papeo, 2020), a schematic relation of support would consist in an empty object file on top of another empty object file, and so on. Similar representations, with object files possibly enriched with thematic roles, may account for the representation of many relational events in infancy (Leslie & Keeble, 1987; Rochat, Striano, & Morgan, 2004; Tatone, Geraci, & Csibra, 2015).

We recently provided direct evidence supporting the proposal that prelexical infants lack discrete representations for abstract relations (Hochmann, 2022). We showed that infants can represent the relation same in a format that is abstract, as it can generalize to novel instances of the relation. However those representations are limited to four same individuals, suggesting that the format of infants’ representations is not something like $S(A,B)$, where A and B would be object representations and S the representation of the relation between those objects, but rather $(X X)$, where X is a variable for an object (see Hochmann, 2022, for the full argumentation). The repetition of the variable carries the relational content *same*, but only symbols for objects are explicitly represented. This view is reinforced by the systematic failure of young children and other animal species in the relational match-to-sample task, where they need to match pairs of the same or different images (e.g., matching square–square to circle–circle and square–star to moon–triangle). If infants and young children possessed discrete symbols S and D for the relations same and different, they should activate S for both square–square and circle–circle, and D for both square–star and moon–triangle, and easily match S to S or D to D . Instead children fail until the age of 4, and only succeed when actively using the words “same” and “different” (Hochmann et al., 2017). Likewise, chimpanzees (and other animal species) fail the relational match-to-sample task, unless they previously acquired external unitary symbols that refer to the relations same and different (Premack, 1983; Thompson, Oden, & Boysen, 1997). These observations highlight a discontinuity along human development. They put forward the hypothesis that relations are initially represented in mental models, and that discrete representations of relations are related to the acquisition of words for

those relations. The discrete symbols for abstract relations are possibly no other than the words that refer to those relations.

Finally, even granting infants the capacity to solve the disjunctive syllogism (Cesana-Arlotti et al., 2018) or to compute negation (Hochmann & Toro, 2021), more experimental work is necessary to describe the format of the representations that permit those performances. Although discrete logical operators could account for these data, other hypotheses are still on the table, including among others, probabilistic representations and inhibitory mechanisms.

In conclusion, the LoT hypothesis is a hypothesis about the format of mental representations. Despite the appeal of a unifying account of cognition and perception, from infancy to adulthood, from bees to humans, discontinuities in the format of thoughts deserve to be studied and highlighted. Quilty-Dunn et al. provide a framework to think about these issues in infants – as well as in nonhuman animals – and develop experimental approaches to decide whether each LoT property is present or absent in infancy, whether infants indeed possess a propositional LoT, or whether they still need to acquire some of the pieces before they can fully play the game.

Financial support. This work was supported by the Agence Nationale pour la Recherche grant ANR-16-CE28-0006 TACTIC and the collaborative McDonnell Foundation Grant 220020449.

Competing interest. None.

References

- Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, 14(3), 110–118.
- Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science (New York, N.Y.)*, 359(6381), 1263–1266.
- Denison, S., & Xu, F. (2010). Twelve- to 14-month-old infants can predict single-event probability with large set sizes. *Developmental Science*, 13(5), 798–803.
- Engelmann, J. M., Haux, L. M., Völter, C., Schleihauf, H., Call, J., Rakoczy, H., & Herrmann, E. (2022). Do chimpanzees reason logically?. *Child Development*. <https://doi.org/10.1111/cdev.13861>
- Goupil, N., Papeo, L., & Hochmann, J. R. (2022). Visual perception grounding of social cognition in preverbal infants. *Infancy*, 27(2), 210–231.
- Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450(7169), 557–559.
- Hochmann, J.-R. (2022). Representations of abstract relations in infancy. *Open Mind*, 6, 291–310. https://doi.org/10.1162/opmi_a_00068
- Hochmann, J.-R., & Papeo, L. (2021). How can it be both abstract and perceptual? Comment on Hafri, A., & Firestone, C. (2021), The perception of relations, *Trends in Cognitive Sciences*. <https://psyarxiv.com/hm49p>
- Hochmann, J. R., & Toro, J. M. (2021). Negative mental representations in infancy. *Cognition*, 213, 104599.
- Hochmann, J. R., Tuerk, A. S., Sanborn, S., Zhu, R., Long, R., Dempster, M., & Carey, S. (2017). Children's representation of abstract relations in relational/array match-to-sample tasks. *Cognitive Psychology*, 99, 17–43.
- Kominsky, J. F., & Scholl, B. J. (2020). Retinotopic adaptation reveals distinct categories of causal perception. *Cognition*, 203, 104339.
- Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science (New York, N.Y.)*, 354(6308), 110–114.
- Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality?. *Cognition*, 25(3), 265–288.
- Papeo, L. (2020). Twos in human visual perception. *Cortex*, 132, 473–478.
- Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences of the United States of America*, 110(41), E3965–E3972.
- Premack, D. (1983). The codes of man and beasts. *Behavioral and Brain Sciences*, 6(1), 125–136.
- Rochat, P., Striano, T., & Morgan, R. (2004). Who is doing what to whom? Young infants' developing sense of social causality in animated displays. *Perception*, 33(3), 355–369.
- Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science (New York, N.Y.)*, 348(6230), 91–94.
- Tatone, D., Geraci, A., & Csibra, G. (2015). Giving and taking: Representational building blocks of active resource-transfer events in human infants. *Cognition*, 137, 47–62.
- Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences of the United States of America*, 104(48), 19156–19159.
- Thompson, R. K., Oden, D. L., & Boysen, S. T. (1997). Language-naïve chimpanzees (*Pan troglodytes*) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes*, 23(1), 31.

Animal thought exceeds language-of-thought

Angelica Kaufmann  and Albert Newen 

Institut für Philosophie II, Ruhr-Universität Bochum, Bochum, Germany
angelica.kaufmann@gmail.com
albert.newen@rub.de
www.angelicakaufmann.com
<https://www.pe.ruhr-uni-bochum.de/philosophie/ii/newen/kontakt.html.en>

doi:10.1017/S0140525X23002017, e279

Abstract

Quilty-Dunn et al. claim that all complex infant and animal reasoning implicate language-of-thought hypothesis (LOTH)-like structures. We agree with the authors that the mental life of animals can be explained in representationalist terms, but we disagree with their idea that the complexity of mental representations is best explained by appealing to abstract concepts, and instead, we explain that it doesn't need to.

Quilty-Dunn et al. claim that “complex infant and animal reasoning [...] all implicate LOT-like structures” (target article, long abstract). The authors explain that recent evidence in comparative psychology shows that “The use of abstract content in physical reasoning is arguably present throughout the animal kingdom” (target article, sect. 5.1, para. 4), and, they say, these findings are compatible with their proposal that language-of-thought hypothesis (LOTH)-based accounts have the potential to explain all animal cognition. We will comment on the authors' claim, given the authors' interpretation of the findings they choose to exemplify in comparative psychology and given their account of animal reasoning and the nature of mental representations they take it to involve.

We agree with the authors that the mental life of animals can be explained by referring to various formats and architectures; we even agree that representational approaches are well suited to stand out as flexible and useful explanatory tools. However, we suggest that from accepting that representationalist accounts of the mind are suitable to explain animal minds, to claiming that mental representations are sentence-like in nonhuman animals, is yet a big step. We want to comment on the authors' proposal that LOTH accounts must hold to six core properties, and in particular, we question the sixth property: Abstract conceptual content. We do not deny that complex cognitive processes take place in much animal mental life, on the contrary. But we disagree

that such complexity is best explained by appealing to abstract concepts.

According to LOTH, the language-of-thought hypothesis, mental representations are formatted like sentences (Fodor, 1975, 1987). A LOTH, language-of-thought hypothesis for animals, has since been discussed (see Beck, 2017), also on the ground that LOTH can be split up into strong-LOT, that understand the compositionality of representations as involving the mechanisms proper of natural languages, and weak-LOT, that maintains the compositionality of representations (Camp, 2007, 2009). To date, there is no direct evidence for LOTH, but there is evidence that some animal mental representations are not sentence-like (see work on analogue magnitudes in Beck, 2015, 2017).

We agree with Fitch's (2020) mentalistic, yet physicalist, perspective that a concept is simply "a nonlinguistic psychological representation of a class of entities in the world." Abstract concepts, in particular, refer to abstract entities, that is, those entities that are at least not directly perceptually available. We argue that animals may not need abstract concepts to engage in complex reasoning. Their mental life is based on representational mechanisms that need not involve abstract concepts. We argue that animals can realize complex activities relying on "minimal beliefs" without involving abstract concepts. Those consist of informational states that are sufficiently decoupled from motivational states, which allows for informational and motivational states to be combinable and organizable. In addition, these representations need to interact with epistemic dispositions to allow for the acquisition of novel information, for their (perception-based) categorization, and for constant updates (Newen & Starzak, 2020). We will put these minimal beliefs at work through an example from primate ethology: Orangutans' long calls (Askew & Morrogh-Bernard, 2016; Lameira & Call, 2018; Spillman et al., 2015; van Schaik, Damerius, & Isler, 2013). These long calls not only lack any involvement of abstract concepts but, as we will elaborate, there are also neither concrete constituents nor a predicate-argument structure involved in them.

An observational study by van Schaik et al. (2013) examined the extent to which the direction of long calls emitted by flanged male Sumatran orangutans (*Pongo abelii*) and Bornean orangutans (*Pongo pygmaeus wurmbii*) indicated the direction of their future travel. These animals live in a very dense tropical forest and are semi-solitary, thus often out of sight from other members of their population. For this reason, their communicative repertoire is distinctively (though not exclusively) more vocal than that of other apes. Flanged male orangutans use long calls to indicate to female members their future travel direction. These male individuals perform these calls when stationary. And these vocalizations can anticipate the direction of their travel 1 day ahead. In response, females show receptive behaviour by travelling in the direction indicated by the long calls. The study of this communicative strategy focused on three questions: First, testing to what extent the direction in which flanged male Sumatran orangutans give spontaneous long calls predicts their travel direction accurately. Second, if the initial calls are followed by additional spontaneous long calls that indicate the subsequent travel direction with more precision than the old one would if no new call had been given. Third, the extent to which long calls that are given in the evening from the night nest or in its proximity still indicate travel direction during the next day, thus indicating future planning independent of the current motivational state.

The capacity displayed by flanged male orangutans to communicate their future travel directions, and the corresponding ability displayed by the females to be receptive to such communicative intentions, is readily explained through the framework of minimal beliefs.

The long calls communicate spatial and temporal information about future travels. This information is first processed, then stored, and eventually reactivated and integrated at the time of the day they will need to be used to guide the travelling. We suggest that picking up different categories of information (direction = inferred by the orientation of the male performing the call; distance = inferred by the loudness of the call; time = inferred by the intervals between one call and the following ones) about the same event (= travel), then processing, storing, and retrieving them can be managed by combinable and organizable informational states like minimal beliefs and that is important evidence of the representational capacities of orangutans. But this does not imply any LOTH-like structure of the mental representation involved in the long calls. More precisely, we do not need to presuppose any language-like syntactic structure, and, specifically, no subject-predicate structure for the long calls.

Even if the authors suggest understanding all communication and reasoning through language-like structures in a wide sense, to justify compositionality, this use of the LOTH would result inflationary and the hypothesis itself would lose its explanatory power. Thus, the LOTH still is not the key explanatory framework to understanding complex cognition in nonhuman animals.

Acknowledgments. We thank the editors of *Behavioural and Brain Sciences* for their support and editorial supervision.

Financial support. This research was supported by the DFG-project (NE 576/14-1) "The structure and development of understanding actions and reasons" (Gefördert durch die Deutsche Forschungsgemeinschaft [DFG] – Projekt NE 576/14-1).

Competing interest. None.

References

- Askew, J. A., & Morrogh-Bernard, H. C. (2016). Acoustic characteristics of long calls produced by male orangutans (*Pongo pygmaeus wurmbii*): Advertising individual identity, context, and travel direction. *Folia Primatologica*, 87, 305–319. doi:10.1159/000452304
- Beck, J. (2015). Analogue magnitude representations: A philosophical introduction. *British Journal for the Philosophy of Science*, 66(4), 829–855.
- Beck, J. (2017). Do nonhuman animals have a language of thought? In K. Andrews & J. Beck (Eds.), *The Routledge handbook of philosophy of animal minds* (pp. 46–55). Routledge/Taylor & Francis Group. <https://doi.org/10.4324/9781315742250-5>
- Camp, E. (2007). Thinking with maps. *Philosophical Perspectives*, 21, 145–182.
- Camp, E. (2009). A language of baboon thought? In R. Lurz (Ed.), *The philosophy of animal minds* (pp. 108–127). Cambridge University Press. doi:10.1017/CBO9780511819001.007
- Fitch, W. T. (2020). Animal cognition and the evolution of human language: Why we cannot focus solely on communication. *Philosophical Transactions of the Royal Society B*, 375(1789), 20190046. <http://doi.org/10.1098/rstb.2019.0046>
- Fodor, J. A. (1975). *Language of thought*. Harvard University Press.
- Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*, M. A. Boden (Ed.). MIT Press
- Lameira, A. R., & Call, J. (2018). Time-space-displaced responses in the orangutan vocal system. *Science Advances*, 4, 11. doi:10.1126/sciadv.aau3401
- Newen, A., & Starzak, T. (2020). How to ascribe beliefs to animals. *Mind and Language*, 37(1), 3–21.
- Spillman, B., van Noordwijk, M. A., Willems, E. P., Mitra Setia, T., Wipfli, U., & van Schaik, C. P. (2015). Validation of an acoustic location system to monitor Bornean orangutan (*Pongo pygmaeus wurmbii*) long calls. *American Journal of Primatology*, 77, 767–776. doi:10.1002/ajp.22398
- van Schaik, Damerius, & Isler. (2013). Wild orangutan males plan and communicate their travel direction one day in advance. *PLoS ONE*, 8(9), e74896. <https://doi.org/10.1371/journal.pone.0074896>

The language-of-thought as a working hypothesis for developmental cognitive science

Melissa M. Kibbe 

Department of Psychological & Brain Sciences, Boston University, Boston, MA, USA

kibbe@bu.edu

<https://www.bu.edu/cdl/developing-minds-lab/>

doi:10.1017/S0140525X23002030, e280

Abstract

A science of prelinguistic infant cognition must take seriously the language-of-thought (LoT) hypothesis. I show how the LoT framework enables us to identify the representational and computational capacities of infant minds and the developmental factors that act on these capacities, and explain how Quilty-Dunn et al.'s take on LoT has important upshots for developmental theory-building.

The language-of-thought (LoT) framework enables us to formulate testable hypotheses about the representational formats, computational structures, and expressive power of infants' thoughts before they can effectively communicate via signed or spoken language. Quilty-Dunn et al.'s take on the LoT hypothesis provides a highly generative framework for operationalizing the relevant units in this hypothesis space. Here, I explain why cognitive developmentalists should embrace such an LoT approach.

LoT-framed hypotheses can already be found throughout foundational studies in infant cognitive science, although they are not necessarily explicitly formulated as such. They can be found in our attempts to understand the early formats of representations of objects, number, space, and agents (e.g., Baillargeon, 2004; Feigenson, Dehaene, & Spelke, 2004; Kibbe, 2015; Leslie, 1994; Spelke & Kinzler, 2007; Vasilyeva & Lourenco, 2012), learning and reasoning (e.g., Denison & Xu, 2019; Rabagliati, Ferguson, & Lew-Williams, 2019), and social cognition (e.g., Kushnir, 2022; Leslie, 1987), to name just a few examples. There is also a growing literature on infants' capacity for combinatorial thought (e.g., Cesana-Arlotti et al., 2018; Piantadosi, Palmeri, & Aslin, 2018). Although there is insufficient space here to engage with all of the nuances of these domains, I aim to draw a throughline: Across these domains infant researchers ask, what are the basic representational units of infants' mental lives? How do infants manipulate these representations in the absence of continued perceptual input? How might infants make new connections between representations, learn new concepts, or think new thoughts? To what extent might these early capacities form the basis of more complex thought or the acquisition of new knowledge domains (like physics or algebra)? Answering these questions requires formulating hypotheses around LoTs.

As a case study, consider the research on infants' capacity for "arithmetic." In a now classic paper, Wynn (1992) found that infants who were shown objects hidden one at a time behind an occluder were able to represent the total quantity of objects hidden. Wynn (1992) suggested that infants' success was evidence

that they grasp the numerical relationship between the inputs and outputs of an addition computation. This is a provocative suggestion – that infants have an LoT-like capacity for combinatorial thought over numerical representations – and the LoT framework allows us to set up testable conditions under which such capacities might be evidenced. Infants could be doing something that resembles the *expressive power of arithmetic*: Their representations of the objects could be formatted in a way that allows those representations to be used as operands in a mental function specifying arithmetic relations (i.e., $f([\text{object } a], [\text{object } b]) = a + b = c$) – consistent with an LoT. Or, infants could be doing something that is decidedly not LoT-like at all: For example, they could track the two sequentially hidden individual objects via separately deployed attentional indexes resulting in a representation of $[\text{index}, \text{index}]$, and the total quantity is represented only implicitly by the number of indexes deployed. In fact, we do not yet know which of these potential explanations (if either) underlies infants' behavior in Wynn (1992) (see Cheng & Kibbe, 2023, for related discussion). Identifying to what extent infants' early capacities have a computational structure or combinatorial capacity similar to formal arithmetic is particularly important because we care not only about what's going on in the infant mind, but also about whether formal numerical knowledge can emerge from infants' early capacities (see, e.g., Carey, 2009). This is just one example, but there are many such LoT hypotheses out there just waiting to be tested.

For developmentalists who may be hesitant to explicitly formulate LoT hypotheses about infant cognition, Quilty-Dunn et al.'s approach to LoT has major advantages. It does not commit infant researchers to a single format for an LoT, into which disparate evidence must be proverbially crammed. Their approach also does not commit us to LoTs with full, recursive, natural-language-like expressive grammar, which would be difficult to square with infants' apparent capacities in a variety of domains. And their approach does not commit developmentalists to Fodor's radical nativism, which in the past has (somewhat unfairly, I would argue) marked the LoT hypothesis as incommensurate with development.

In fact, an LoT approach is developmental. Taking an LoT approach to infant cognition is compatible with efforts to understand how motor development, neurobiological development, and/or individual differences related to cognitive, emotional, educational, economic, or sociocultural factors may shape cognition in infancy and beyond. Indeed, it allows us to formulate hypotheses to identify potential computational structures over which these factors may operate across early development. It allows for the possibility of differential developmental trajectories for LoTs across domains and across infancy.

Taking an LoT approach also does not require that infants' early computational capacities must be quarantined from experience or learning (e.g., I think positing that LoT structures may be identifiable in infancy is compatible with a rational constructivist approach; see Xu, 2019). Nor does it require us to find evidence for LoT(s) in every aspect of early cognition. There are plenty of instances in which iconic, noncombinatorial representations provide the best explanation for infants' (and, indeed, adults') behavior in some domain. Instead, it requires infant researchers to take seriously the possibility that infants can think before they articulate language, and to identify cases of such expressive capacities, and their functional utility for the developing mind.

Importantly, taking an LoT approach also does not entail that there is some developmental hierarchy in the expressive power of

LoTs, with infants on the bottom and adults at the top. Infant researchers can and should formulate hypotheses around both continuity and change in the computational capacities of minds across domains and across development.

Quilty-Dunn et al. lay out the case for LoT-like structures in the mind, and a roadmap for how to go about looking for them. If human cognition includes LoTs in its fully developed state, then we need to formulate our hypotheses about the origins and development of human cognition within an LoT framework. Cognitive developmentalists should embrace this approach.

Acknowledgment. The author is grateful to Derek E. Anderson for helpful discussions.


Financial support. This work was supported by the National Science Foundation (BCS 1844155).

Competing interest. None.

References

- Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, 13(3), 89–94.
- Carey, S. (2009). Where our number concepts come from. *The Journal of Philosophy*, 106(4), 220–254.
- Cesana-Arlotti, N., Martin, A., Téglás, E., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science (New York, N.Y.)*, 359(6381), 1263–1266.
- Cheng, C., & Kibbe, M. M. (2023). Is non-symbolic arithmetic truly “arithmetic”? Examining the computational capacity of the approximate number system in young children. *Cognition* 47, e13299.
- Denison, S., & Xu, F. (2019). Infant statisticians: The origins of reasoning under uncertainty. *Perspectives on Psychological Science*, 14(4), 499–509.
- Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, 8(7), 307–314.
- Kibbe, M. M. (2015). Varieties of visual working memory representation in infancy and beyond. *Current Directions in Psychological Science*, 24(6), 433–439.
- Kushnir, T. (2022). Imagination and social cognition in childhood. *Wiley Interdisciplinary Reviews: Cognitive Science*, 13(4), e1603.
- Leslie, A. M. (1987). Pretense and representation: The origins of “theory of mind”. *Psychological Review*, 94(4), 412.
- Leslie, A. M. (1994). ToMM, ToBy, and agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 119–148). Cambridge University Press.
- Piantadosi, S. T., Palmeri, H., & Aslin, R. (2018). Limits on composition of conceptual operations in 9-month-olds. *Infancy*, 23(3), 310–324.
- Rabagliati, H., Ferguson, B., & Lew-Williams, C. (2019). The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental Science*, 22(1), e12704.
- Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, 10(1), 89–96.
- Vasilyeva, M., & Lourenco, S. F. (2012). Development of spatial cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(3), 349–362.
- Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, 358, 749–750.
- Xu, F. (2019). Towards a rational constructivist theory of cognitive development. *Psychological Review*, 126(6), 841.

Is language-of-thought the best game in the town we live?

Gary Lupyan 

University of Wisconsin-Madison, Madison, WI, USA
lupyan@wisc.edu; <http://sapir.psych.wisc.edu>

doi:10.1017/S0140525X23001814, e281

Abstract

There are towns in which language-of-thought (LoT) is the best game. But do we live in one? I go through three properties that characterize the LoT hypothesis: Discrete constituents, role-filler independence, and logical operators, and argue that in each case predictions from the LoT hypothesis are a poor fit to actual human cognition. As a hypothesis of what human cognition ought to be like, LoT departs from empirical reality.

The effort by Quilty-Dunn et al. to evaluate the language-of-thought hypothesis (LoTH) in light of what has been learned since Fodor's original formulation is commendable. But although it is possible to interpret some behaviors as being compatible with LoT, LoT remains a poor way to understand human cognition. If the target article is the “strongest article-sized empirical case for LoTH” (target article, sect. 1, para. 4), the case of LoT is rather weak.

Let us examine three properties of LoTH. For each, I will consider what we might expect if the property actually holds of human cognition and what we instead tend to find. The reasoning applies to the remaining three properties, but space prohibits further explication.

Discrete constituents: It is true that the English sentence “That is a pink square object” can be decomposed into constituents like “pink” and “square” that can be plugged into other sentences to convey something of the same meaning. Two problems. First, the authors are making a case for discrete constituents of *thought*, but support their core argument with examples from language. It is one thing to show that language has certain properties. It is quite another to show that these properties characterize *thoughts* (Lupyan, 2016; Mahowald et al., 2023; Malt & Majid, 2013; Malt et al., 2015). Supporting the latter would require showing that underlying our language use are discrete concepts (if one holds onto Fodor's extreme nativism, these concepts are also innate – an even higher bar). Evidence against such a view is too lengthy to review here (Levinson, 1997; Lupyan & Zettersten, 2021; Malt & Majid, 2013), but consider the fuzziness and context-dependence of even the easiest-to-define concepts like ODD, EVEN, and TRIANGLE (Lupyan, 2013, 2015). Second, even language may not be as discrete as is often assumed. To us, literate English-speaking scholars with a habit of reflecting on language as an external artifact, the idea that it is composed of discrete parts may seem self-evident. But this may speak more to what it *can* be than what it typically *is*. For example, literate, but not illiterate children can count words in a spoken sentence (Matute et al., 2012; Olson, 2002) – a surprising result if natural language simply maps onto discrete constituents of thought.

Role-filler independence: John is the agent of “John loves Mary” in the same way that Mary is the agent of “Mary loves John.” Does this mean that role-filler independence is a characteristic property of our thoughts? Even if it were, this does not mean that role-filler independence is a core property of (nonlinguistic) cognition. But never mind that. *Agent* together with *patient* does indeed turn out to be perhaps the strongest example of role-filler independence (Rissman & Majid, 2019). However, Rissman and Majid go on to argue that evidence for the abstract nature of other seemingly basic roles like *instrument* and *goal* is rather mixed. Even for *agent*, role-filler independence is more subtle than it seems. In a nonlinguistic task requiring participants to categorize based on agent/patient relationships, a sizable minority (~40%) failed to induce it in the allotted time (Rissman & Lupyan, 2022). Those who did, generalized agency according to how similar the

test items were to the items they saw at training as well as to the test item's similarity to agent prototypes (Dowty, 1991). It seems that not all agents are equally good agents, a surprising result if there is true role-filler independence.

The authors correctly point out that connectionist models “simulate compositionality, but fail to preserve identity of the original representational elements” (target article, sect. 2, para. 7). The authors do not consider the possibility that human compositionality may be simulated as well (Dekker, Otto, & Summerfield, 2022; Lahav, 1989).

Lastly, *logical operators* such as AND, IF, and OR are a “hallmark of LoT architectures” (target article, sect. 2, para. 10). Yet children under the age of about five have a notoriously difficult time learning categories based on even the simplest logical rules (Rabi, Miles, & Minda, 2015; Rabi & Minda, 2014). Adults are better (and certainly better than other animals!), but arguably rule-based reasoning is far more difficult than it should be if such logical operators actually underlie much of our perception and reasoning (Goldwater, Don, Krusche, & Livesey, 2018; Lupyan, 2013; Mercier & Sperber, 2017).

It is true that at least for stimuli composed of easy-to-verbalize and recombine features such as circles and triangles of various colors used by Piantadosi, Tenenbaum, and Goodman (2016) adults can do well, showing patterns of behavior well-explained by the use of logical operators. However, such behavior is fragile in ways unexpected if these operators underlie our everyday cognition. Formally simple operations like XOR are notoriously difficult for people (Shepard, Hovland, & Jenkins, 1961). Even on simple rules like IF A, performance strongly depends on factors like verbal nameability of the constituents (Zettersten & Lupyan, 2020).

Ironically, Piantadosi, cited in support of hard-coded logical connectives (Piantadosi et al., 2016) was explicit that their data concern adults (“our results are not about children,” p. 22) making the claim that logical operators underlie our core cognitive processes suspect. He later went on to argue that “primitives” like AND and OR need not in fact be primitives and can be learned (Piantadosi, 2021). I would add that such learning may be supported in part by natural language (Lupyan & Bergen, 2016).

To be fair, not all the evidence the authors use in support of the LoTH is linguistic. A considerable weight is placed on the construct of object files that are somehow meant to explain perception in terms of LoTH. Although object files may be a useful construct for understanding certain perceptual generalizations, there is good reason why research in perception treats visual representations as analog/iconic representations (Block, forthcoming).

In a town inhabited by highly educated people with a Western philosophical bent, LoTH is a sensible starting point in thinking about how cognition works. In towns inhabited by the rest of us, it is a curious game that some learn to play. The most fun games are often those that transport us to imagined worlds. The world of the LoT hypothesis is likely one of these.

Financial support. This study was supported by NSF-PAC 2020969.

Competing interest. None.

References

- Block, N. (forthcoming). Let's get rid of the concept of an object file. In B. McLaughlin & J. Cohen (Eds.), *Contemporary debates in philosophy of mind* (2nd ed., pp. 494–516). Wiley. <https://philarchive.org/rec/BLOLGR>
- Dekker, R. B., Otto, F., & Summerfield, C. (2022). Curriculum learning for human compositional generalization. *Proceedings of the National Academy of Sciences of the United States of America*, 119(41), e2205582119. <https://doi.org/10.1073/pnas.2205582119>

- Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3), 547–619. <https://doi.org/10.2307/415037>
- Goldwater, M. B., Don, H. J., Krusche, M. J. F., & Livesey, E. J. (2018). Relational discovery in category learning. *Journal of Experimental Psychology: General*, 147(1), 1–35. <https://doi.org/10.1037/xge0000387>
- Lahav, R. (1989). Against compositionality: The case of adjectives. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 57(3), 261–279.
- Levinson, S. C. (1997). From outer to inner space: Linguistic categories and non-linguistic thinking. In J. Nuyts & E. Pederson (Eds.), *Language and conceptualization* (pp. 13–45). Cambridge University Press.
- Lupyan, G. (2013). The difficulties of executing simple algorithms: Why brains make mistakes computers don't. *Cognition*, 129(3), 615–636. <https://doi.org/10.1016/j.cognition.2013.08.015>
- Lupyan, G. (2015). The paradox of the universal triangle: Concepts, language, and prototypes. *Quarterly Journal of Experimental Psychology*, 70(3), 389–412. <https://doi.org/10.1080/17470218.2015.1130730>
- Lupyan, G. (2016). The centrality of language in human cognition. *Language Learning*, 66(3), 516–553. <https://doi.org/10.1111/lang.12155>
- Lupyan, G., & Bergen, B. (2016). How language programs the mind. *Topics in Cognitive Science*, 8(2), 408–424. <https://doi.org/10.1111/tops.12155>
- Lupyan, G., & Zettersten, M. (2021). Does vocabulary help structure the mind? In M. D. Sera & M. A. Koenig (Eds.), *Minnesota symposia on child psychology* (pp. 160–199). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119684527.ch6>
- Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). *Dissociating language and thought in large language models: A cognitive perspective*. arXiv: 2301.06627. <https://doi.org/10.48550/arXiv.2301.06627>
- Malt, B. C., Gennari, S., Imai, M., Ameel, E., Saji, N., & Majid, A. (2015). Where are the concepts? What words can and can't reveal. In E. Margolis & S. Laurence (Eds.), *Concepts: New directions* (pp. 291–326). MIT Press.
- Malt, B. C., & Majid, A. (2013). How thought is mapped into words. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(6), 583–597. <https://doi.org/10.1002/wics.1251>
- Matute, E., Montiel, T., Pinto, N., Rosselli, M., Ardila, A., & Zarabozo, D. (2012). Comparing cognitive performance in illiterate and literate children. *International Review of Education*, 58(1), 109–127. <https://doi.org/10.1007/s11159-012-9273-9>
- Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.
- Olson, D. R. (2002). What writing does to the mind. In E. Amsel & J. P. Byrnes (Eds.), *Language, literacy, and cognitive development* (pp. 153–165). Erlbaum.
- Piantadosi, S. T. (2021). The computational origin of representation. *Minds and Machines*, 31(1), 1–58.
- Piantadosi, S. T., Tenenbaum, J., & Goodman, N. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, 123(4), 392–424.
- Rabi, R., Miles, S. J., & Minda, J. P. (2015). Learning categories via rules and similarity: Comparing adults and children. *Journal of Experimental Child Psychology*, 131, 149–169. <https://doi.org/10.1016/j.jecp.2014.10.007>
- Rabi, R., & Minda, J. P. (2014). Rule-based category learning in children: The role of age and executive functioning. *PLoS ONE*, 9(1), e85316. <https://doi.org/10.1371/journal.pone.0085316>
- Rissman, L., & Lupyan, G. (2022). A dissociation between conceptual prominence and explicit category learning: Evidence from agent and patient event roles. *Journal of Experimental Psychology: General*, 151(7), 1707–1732. <https://doi.org/10.1037/xge0001146>
- Rissman, L., & Majid, A. (2019). Thematic roles: Core knowledge or linguistic construct? *Psychonomic Bulletin & Review*, 26(6), 1850–1869. <https://doi.org/10.3758/s13423-019-01634-5>
- Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, 75(13), 1–42. <https://doi.org/10.1037/h0093825>
- Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition*, 196, 104135. <https://doi.org/10.1016/j.cognition.2019.104135>

Stop me if you've heard this one before: The Chomskyan hammer and the Skinnerian nail

Alex Madva 

California State Polytechnic University, Pomona, CA, USA
alexmadva@gmail.com
<https://alexmadva.com/>

doi:10.1017/S0140525X23001851, e282

Abstract

The target article signal boosts important ongoing work across the cognitive sciences. However, its theoretical claims, generative value, and purported contributions are – where not simply restatements of arguments extensively explored elsewhere – imprecise, noncommittal, and underdeveloped to a degree that makes them difficult to evaluate. The article’s apparent force results from engaging with straw rather than steel opponents.

Batman: Then why do you want to kill me?

Joker: Kill you? I don’t want to kill you. What would I do without you? Go back to ripping off mob dealers? No, no, no. No, you... you complete me. (Nolan, 2008)

For many a hammer, everything is a nail. For many a philosopher of mind, everything is a chance to rehearse Kant’s criticism of Hume, Chomsky’s criticism of Skinner, and Fodor’s criticism of every empiricist, holist, or, in the pages of *BBS* (1985), relativist who rubbed him the wrong way. Thus the target article repeats, again and again across different domains, a well-worn argumentative maneuver in psychology and philosophy: “this mental phenomenon you’re trying to explain in terms of a simpler process – be it associations, model-free learning, neural nets, or icons – must instead be explained by a more complex process, which performs language-like computations.”

The theoretical payoff of this selective tour through case studies is remarkably modest. For Quilty-Dunn et al. do not deny that the simpler, nonlinguistic processes exist and have real effects. They deny that the nonlinguistic processes explain *everything*. Repurposing the old joke, we’ve established what kind of theorist you are – a pluralist – and now we’re just haggling over the details. The haggling, in this case, recalls trench warfare. On some fronts (like artificial intelligence), neural nets make stunning advances, even as the language-of-thought hypothesis (LoTH) plants its flag on other patches of cognitive terrain hitherto claimed by non-linguistic theories. The broader import of this unsystematic assemblage of localized skirmishes is unclear. Is LoTH “the best game in town,” or one game among others, which the mind perhaps plays somewhat more often than some think?

The authors deny that so-called “system 1” (target article, sect. 6, note 11) is *purely* associative. It’s true that associative interpretations of implicit bias continue to hold sway in pop-psych discourse, but hardly anyone paying attention to what the authors rightly call a “near-deluge” (target article, sect. 6.2, para. 5) of research on propositional effects in implicit social cognition continues to defend the extreme associationist views targeted by the authors. Many of us never did (Brownstein & Madva, 2012; Del Pinal, Madva, & Reuter, 2017; Gawronski & Bodenhausen, 2006, pp. 706–707; Madva, 2016, p. 2681, 2019; see also Brownstein, 2018, Chs. 2–3). This is not to say all our predictions panned out, but to question the ease with which other pluralist approaches are pigeonholed into the dreaded empiricist/associationist/behaviorist position in these recurring debates (e.g., Kurdi, Morris, & Cushman, 2022b, p. 3). Indeed, according to Mandelbaum (2022, sect. 8), we represent “a revival of associationist theories in philosophy,” citing a paper that is explicitly orthogonal to the association–proposition debate (Madva & Brownstein, 2018, sect. 6.1; see also Kurdi, Mann, Charlesworth, & Banaji, 2019; Phills, Hahn, & Gawronski, 2020). With apologies to Voltaire, one senses that if modern-day associationists did not exist, modern-day Fodorians would have to reinvent them. With apologies to Taylor Swift, I would very much like to be excluded from this narrative.

In any case, the downfall of pure-associative models has not occasioned the uncontested reign of propositional alternatives. Leading propositional theorists continue to uncover effects more naturally explained by nonpropositional processes, or at least uneasily assimilated into prevailing propositional theories (e.g., Van Dessel, De Houwer, Gast, Roets, & Smith, 2020; see also Byrd, 2021). As a recent meta-analysis by Kurdi, Morehouse, and Dunham (2023, p. 1) explains, no current theory is well-poised to predict and explain the disorienting array of findings, and the time for “existence proof demonstrations” of propositional effects has passed. Yet in lieu of synthesizing the disarray, the target article consists in just such a grab bag of existence proofs, trumpeting all and only recent successes for propositional approaches – while ignoring evidence of their shortcomings and boundary conditions, and deferring long-standing concerns about how LoTs are implemented in the brain and integrated with other processes.

The authors nevertheless advertise LoTH’s “unificatory power” (target article, sect. 1, para. 7), specifically its provision of a *lingua franca* mediating between psychological domains (perception, higher-order thinking, so-called “system 1,” etc.). But if each of these domains involves proprietary LoTs *and who knows how many other* representational formats, the question still remains how these diverse representational formats interact with each other (within each psychological domain, rather than between domains). If non-LoTs interface with LoTs after all, what explanatory traction is gained by noting how LoTs pop up in lots of distinct psychological locales? And if a thousand other representational formats are abloom across the mind (target article, sect. 2, para. 2), why couldn’t some of them mediate between domains, too? What we have here is not unification but proliferation, not explanation but more to explain.

No doubt the authors would cite their six core LoT properties as significant theoretical contributions. But the conceptual and causal interrelations of these properties (which are invoked in seemingly random combinations from one case study to the next) are muddled at best. Do they represent a homeostatic property cluster, as the authors claim, or are they tied more tightly together? The authors stress that representations involving discrete constituents need not be structured like sentences, but they “usually interpret” sentence-like representations “as requiring” discrete constituents (target article, sect. 2, para. 8). They then grant that successive properties on their list necessitate others, for example, predicate–argument structures and logical operators “requiring role-filler independence” (target article, sect. 2, para. 9). To the extent that property B *requires* property A, it is completely trivial to predict that A will show up wherever B does, and only slightly less trivial to predict that B will appear alongside A above chance. The mere prediction that properties “should tend to cooccur” (target article, sect. 2, para. 12) is weak, vague, and unconstrained, allowing theorists to underscore cooccurrences and ignore (or explain away) noncooccurrences. We “usually require” fewer degrees of freedom from our theoretical frameworks. We are also compelled to ask whether the six properties offer anything substantively novel or illuminating, or simply stick new labels on the analytic entailments contained in the original LoT view.

The target article at times positions itself as a lone voice of logic in an associationist wilderness, fighting the good fight for a nearly forgotten rationalist cause while flanked on all sides by zombie empiricisms that refuse to stay dead. Yet the article’s principal value consists in signal-boosting others’ important ongoing work. The question, then, is what it would mean to take up the authors’ proposals over and above what the exemplary researchers being cited are already doing.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

- Brownstein, M. (2018). *The implicit mind: Cognitive architecture, the self, and ethics*. Oxford University Press.
- Brownstein, M., & Madva, A. (2012). The normativity of automaticity. *Mind & Language*, 27(4), 410–434. <https://doi.org/10.1111/j.1468-0017.2012.01450.x>
- Byrd, N. (2021). What we can (and can't) infer about implicit bias from debiasing experiments. *Synthese*, 198(2), 1427–1455. <https://doi.org/10.1007/s11229-019-02128-6>
- Del Pinal, G., Madva, A., & Reuter, K. (2017). Stereotypes, conceptual centrality and gender bias: An empirical investigation. *Ratio*, 30(4), 384–410. <https://doi.org/10.1111/rati.12170>
- Fodor, J. A. (1985). Précis of the modularity of mind. *Behavioral and Brain Sciences*, 8(1), 1–5. <https://doi.org/10.1017/S0140525X0001921X>
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132(5), 692–731. <https://doi.org/10.1037/0033-2909.132.5.692>
- Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences*, 116(13), 5862–5871. <https://doi.org/10.1073/pnas.1820240116>
- Kurdi, B., Morehouse, K. N., & Dunham, Y. (2023). How do explicit and implicit evaluations shift? A preregistered meta-analysis of the effects of co-occurrence and relational information. *Journal of Personality and Social Psychology*, 124(6), 1174–1202. <https://doi.org/10.1037/pspa0000329>
- Kurdi, B., Morris, A., & Cushman, F. A. (2022b). The role of causal structure in implicit evaluation. *Cognition*, 225, 105116. <https://doi.org/10.1016/j.cognition.2022.105116>
- Madva, A. (2016). Why implicit attitudes are (probably) not beliefs. *Synthese*, 193(8), 2659–2684. <https://doi.org/10.1007/s11229-015-0874-2>
- Madva, A. (2019). Social psychology, phenomenology, and the indeterminate content of unreflective racial bias. In E. S. Lee (Ed.), *Race as phenomena: Between phenomenology and philosophy of race* (pp. 87–106). Rowman & Littlefield.
- Madva, A., & Brownstein, M. (2018). Stereotypes, prejudice, and the taxonomy of the implicit social mind. *Nous*, 52(3), 611–644. <https://doi.org/10.1111/nous.12182>
- Mandelbaum, E. (2022). Associationist theories of thought. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (winter). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2022/entries/associationist-thought/>
- Nolan, C. (2008). *The dark knight*. Warner Bros.
- Phills, C. E., Hahn, A., & Gawronski, B. (2020). The bidirectional causal relation between implicit stereotypes and implicit prejudice. *Personality and Social Psychology Bulletin*, 46(9), 1318–1330. <https://doi.org/10.1177/0146167219899234>
- Van Dessel, P., De Houwer, J., Gast, A., Roets, A., & Smith, C. T. (2020). On the effectiveness of approach-avoidance instructions and training for changing evaluations of social groups. *Journal of Personality and Social Psychology*, 119, e1–e14. <https://doi.org/10.1037/pspa0000189>

A language of episodic thought?

Johannes B. Mahr  and Daniel L. Schacter

Department of Psychology, Harvard University, Cambridge, MA, USA
jmahr@fas.harvard.edu;
dls@wjh.harvard.edu

doi:10.1017/S0140525X2300198X, e283

Abstract

We propose that episodic thought (i.e., episodic memory and imagination) is a domain where the language-of-thought hypothesis (LoTH) could be fruitfully applied. On the one hand, LoTH could explain the structure of what is encoded into and retrieved from long-term memory. On the other, LoTH can help make sense of how episodic contents come to play such a large variety of different cognitive roles after they have been retrieved.

Quilty-Dunn et al. convincingly show that language-of-thought hypothesis (LoTH) is alive and kicking in contemporary cognitive science. One domain they do not discuss, however, is episodic memory and imagination (i.e., episodic thought). This is not surprising: Traditionally, episodic thought has been widely viewed in terms of iconic forms of representation. Nonetheless, we believe that episodic thought is rife for being theorized in terms of LoTH. Most importantly, LoTH generates novel perspectives on how humans achieve such remarkable productivity and flexibility when thinking about other places and times.

Recent research on episodic memory and imagination suggests that different kinds of episodic thoughts (past memories, future imaginations, counterfactual imaginations, etc.) are cognitively not individuated through their contents (Addis, 2020, 2018; Mahr, 2020; Schacter et al., 2012); that is, episodic contents are “taxonomically neutral” with respect to their cognitive role as imaginations or memories. For example, Mahr, Greene, and Schacter (2021; see also De Brigard, Gessell, Yang, Stewart, & Marsh, 2020) found that participants’ ability to recall the contents of a previously imagined event only weakly predicts their ability to recall whether this event was about the future or the past. This finding suggests that whether a given episode is taken to represent the past or the future (say) is not determined by what is retrieved from memory (i.e., episodic content) but by processes that occur before or after such retrieval (Mahr, 2020).

With this in mind, we propose that there are two main ways in which LoTH can be cashed out in episodic thought. On the one hand, LoTH can help to conceptualize the structure of episodic contents: What is *encoded* into long-term memory and how these contents are later *retrieved* in the service of the construction of both episodic memories and imaginations. According to the “constructive episodic simulation” hypothesis (Schacter & Addis, 2007), episodic retrieval consists in the flexible recombination of the elements of previously encoded experiences. Although there is good evidence to support this idea (see, e.g., Schacter & Addis, 2020, for a review), it remains unclear what mechanisms allow such flexible recombination of episodic elements in the service of episodic simulation. These processes are most commonly thought of in terms of associative inference (Addis, 2020; Carpenter & Schacter, 2017; Horner & Burgess, 2013) even though – as Quilty-Dunn et al. point out – LoT-style representations like scene-grammars (Vö, 2021), object files (Zimmer & Ecker, 2010), and event files (Hommel, 2004) play a role in structuring the information encoded into long-term memory. Similarly, these representations might play a role in structuring what content is retrieved and how it is composed. The fact that episodic contents could thus exhibit LoT properties – contributing to the flexibility and productivity of episodic simulation – has so far been underexplored. For example, evidence for the influence of “schemas” in episodic encoding and retrieval (Irish & Piguet, 2013; Renoult, Irish, Moscovitch, & Rugg, 2019), which also play a role in episodic simulation of future scenarios (Wynn, van Genugten, Sheldon, & Schacter, 2022), might be understood in this light (e.g., Draschkow, Wolfe, & Vo, 2014; Vö & Wolfe, 2013).

On the other hand, LoTH can help to understand how episodic contents come to play their respective cognitive roles. In the minds of adult humans, episodic contents can fill a variety of different roles – for example, as imaginations of past counterfactuals (De Brigard, Addis, Ford, Schacter, & Giovanello, 2013) or representations of event types (Addis, Pan, Vu, Laiser, & Schacter, 2009). A complete theory of episodic simulation requires an account of how “taxonomically neutral” episodic

contents come to fill these roles. Mahr (2020, 2022; see also Mahr & Csibra, 2018) argued that (in addition to their contents) episodic thoughts have (at least) five discrete constituents: *Temporal orientation* (is this event occurring in the past, present, or future?; Mahr & Schacter, 2022; Mahr et al., 2021), *specificity* (is this a unique occurrence or a type of occurrence?; Addis et al., 2009), *subjectivity* (who is the subject of this experience?; Pillemer, Steiner, Kuwabara, Thomsen, & Svob, 2015), *factuality* (did this/will this really happen?; Johnson & Raye, 1981; Simons, Garrison, & Johnson, 2017), and *mnemicity* (do I know about this through my own experience?; Mahr, in press; Mahr et al., 2023). As a result, episodic thoughts fulfill many of the diagnostic features of LoT-style representations proposed by Quilty-Dunn et al.

First, episodic thoughts exhibit *predicate-argument structure*. Complete episodic thoughts are the result of the predication of episodic contents by compounds of these constituents according to syntactic rules. For example, an episodic memory can be analyzed as *remember*[*past, specific, factual, self (EPISODE)*], where “EPISODE” refers to episodic content. Evidence about the independence of episodic contents and assignments of temporality (Mahr & Schacter, 2022; Mahr et al., 2021) support this idea. Further, one can remember imagining suggesting that some of these predicates can be recursively embedded.

Second, these predicates are *discrete*: Episodic thoughts are the result of the composition of distinct conceptual constituents into an LoT-like “sentence.” Although there is not complete independence between these predicates, the space of possible episodic thoughts includes a large number of such sentences (see Mahr, 2020; Michaelian, 2016). For example, the above might be easily amended to *imagine*[*past, specific, factual, self (EPISODE)*].

Third, several phenomena attest to the fact that this architecture exhibits a large degree of *roll-filler independence* – the same episodic content might fill different cognitive roles. For example, people are able to “recast” memories of past events into the future (e.g., one might imagine a basketball game in the past but also imagine the same game as occurring in the future; Thakral, Yang, Addis, & Schacter, 2021), can be convinced to “disbelieve” their memories (Otgaar, Scoboria, & Mazzoni, 2014), and regularly change their assessment of whether they are remembering or imagining an episode (Loftus & Pickrell, 1995).

Finally, the predicates of episodic thought are *abstract*: Although (say) the pastness or futurity of an episode might commonly go along with different contents, temporal orientation itself cannot be depicted (Mahr, 2020; Matthen, 2010).

The hypothesis that episodic thoughts indeed exhibit these features crucially generates a package of unique predictions (for a first pass at testing role-filler independence, see Mahr et al., 2021; and for discreteness, see Mahr & Schacter, 2022). Even though research on the role of LoT in episodic thought is still in its inception, there are potentially large theoretical payoffs for taking LoTH seriously in this domain.

Financial support. This work was supported by a Walter-Benjamin Fellowship (MA 9499/1-1) by the German Research Foundation (DFG) to J. B. M.

Competing interest. None.

References

Addis, D. R. (2018). Are episodic memories special? On the sameness of remembered and imagined event simulation. *Journal of the Royal Society of New Zealand*, 48(2–3), 64–88.

- Addis, D. R. (2020). Mental time travel? A neurocognitive model of event simulation. *Review of Philosophy and Psychology*, 11(2), 233–259.
- Addis, D. R., Pan, L., Vu, M. A., Laiser, N., & Schacter, D. L. (2009). Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering. *Neuropsychologia*, 47(11), 2222–2238.
- Carpenter, A. C., & Schacter, D. L. (2017). Flexible retrieval: When true inferences produce false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(3), 335.
- De Brigard, F., Addis, D. R., Ford, J. H., Schacter, D. L., & Giovanello, K. S. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia*, 51(12), 2401–2414.
- De Brigard, F., Gessell, B., Yang, B. W., Stewart, G., & Marsh, E. J. (2020). Remembering possible times: Memory for details of past, future, and counterfactual simulations. *Psychology of Consciousness: Theory, Research, and Practice*, 7(4), 331.
- Draschkow, D., Wolfe, J. M., & Vo, M. L. H. (2014). Seek and you shall remember: Scene semantics interact with visual search to build better memories. *Journal of Vision*, 14(8), 10–10.
- Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences*, 8(11), 494–500.
- Horner, A. J., & Burgess, N. (2013). The associative structure of memory for multi-element events. *Journal of Experimental Psychology: General*, 142(4), 1370.
- Irish, M., & Piguet, O. (2013). The pivotal role of semantic memory in remembering the past and imagining the future. *Frontiers in Behavioral Neuroscience*, 7, 27.
- Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, 88(1), 67–85.
- Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals*, 25(12), 720–725.
- Mahr, J. B. (2020). The dimensions of episodic simulation. *Cognition*, 196, 104085.
- Mahr, J. B. (2022). Episodic memory: And what is it for? In A. Sant’Anna, J. McCarroll, & K. Michaelian (Eds.), *Current controversies in philosophy of memory* (pp. 149–166). Routledge.
- Mahr, J. B., & Csibra, G. (2018). Why do we remember? The communicative function of episodic memory. *Behavioral and Brain Sciences*, 41, e1.
- Mahr, J. B., Greene, J. D., & Schacter, D. L. (2021). A long time ago in a galaxy far, far away: How temporal are episodic contents?. *Consciousness and Cognition*, 96, 103224.
- Mahr, J. B. (in press). How to become a memory: Individual and collective aspects of mnemicity. *Topics in Cognitive Science*. <https://doi.org/10.1111/tops.12646>
- Mahr, J. B., & Schacter, D. L. (2022). Mnemicity versus temporality: Distinguishing between components of episodic representations. *Journal of Experimental Psychology: General*, 151(10), 2448–2465.
- Mahr, J. B., Van Bergen, P., Sutton, J., Schacter, D. L., & Heyes, C. (2023). Mnemicity: A cognitive gadget? *Perspectives on Psychological Science*, 17456916221141352.
- Matthen, M. (2010). Is memory preservation? *Philosophical Studies*, 148(1), 3–14.
- Michaelian, K. (2016). *Mental time travel*. MIT Press.
- Otgaar, H., Scoboria, A., & Mazzoni, G. (2014). On the existence and implications of nonbelieved memories. *Current Directions in Psychological Science*, 23(5), 349–354.
- Pillemer, D. B., Steiner, K. L., Kuwabara, K. J., Thomsen, D. K., & Svob, C. (2015). Vicarious memories. *Consciousness and Cognition*, 36, 233–245.
- Renoult, L., Irish, M., Moscovitch, M., & Rugg, M. D. (2019). From knowing to remembering: The semantic-episodic distinction. *Trends in Cognitive Sciences*, 23(12), 1041–1057.
- Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society of London B*, 362, 773–786.
- Schacter, D. L., & Addis, D. R. (2020). Memory and imagination: Perspectives on constructive episodic simulation. In A. Abraham (Ed.), *The Cambridge handbook of the imagination* (pp. 111–131). Cambridge University Press.
- Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: Remembering, imagining, and the brain. *Neuron*, 76, 677–694.
- Simons, J. S., Garrison, J. R., & Johnson, M. K. (2017). Brain mechanisms of reality monitoring. *Trends in Cognitive Sciences*, 21(6), 462–473.
- Thakral, P. P., Yang, A. C., Addis, D. R., & Schacter, D. L. (2021). Divergent thinking and constructing future events: Dissociating old from new ideas. *Memory*, 29(6), 729–743.
- Vö, M. L. H. (2021). The meaning and structure of scenes. *Vision Research*, 181, 10–20.
- Vö, M. L. H., & Wolfe, J. M. (2013). The interplay of episodic and semantic memory in guiding repeated search in scenes. *Cognition*, 126(2), 198–212.
- Wynn, J. S., van den Lely, R. D. I., Sheldon, S., & Schacter, D. L. (2022). Schema-related eye movements support episodic simulation. *Consciousness and Cognition*, 100, 103302.
- Zimmer, H. D., & Ecker, U. K. (2010). Remembering perceptual features unequally bound in object and episodic tokens: Neural mechanisms and their electrophysiological correlates. *Neuroscience & Biobehavioral Reviews*, 34(7), 1066–1079.

Properties of LoTs: The footprints or the bear itself?

Sam Whitman McGrath^{a,b*}, Jacob Russin^{c*},

Ellie Pavlick^{a,b} and Roman Feiman^{a,b} 

^aDepartment of Philosophy, Brown University, Providence, RI, USA;

^bDepartment of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, USA and ^cCenter for Neuroscience, University of California, Davis, Davis, CA, USA

sam_mcgrath1@brown.edu

jlrussin@ucdavis.edu

ellie_pavlick@brown.edu

roman_feiman@brown.edu

<https://jlrussin.github.io/>

<https://cs.brown.edu/people/epavlick/index.html>

<https://sites.brown.edu/btllab/>

doi:10.1017/S0140525X23001863, e284

Abstract

There are two ways to understand any proposed properties of language-of-thoughts (LoTs): As diagnostic or constitutive. We argue that this choice is critical. If candidate properties are diagnostic, their homeostatic clustering requires explanation via an underlying homeostatic mechanism. If constitutive, there is no clustering, only the properties themselves. Whether deep neural networks (DNNs) are alternatives to LoTs or potential implementations turn on this choice.

Quilty-Dunn et al. offer six properties of language-of-thoughts (LoTs). Setting aside whether those are the right ones, all proponents of the language-of-thought hypothesis (LoTH) must specify the status of candidate properties: Are they *diagnostic* or *constitutive*? On the diagnostic view, these properties are indicators. Their presence is evidence for an underlying LoT-like representational format, but the format itself would be a distinct natural kind, causally prior to the properties. On the constitutive view, *what it is* to be an LoT is to exhibit some (or all) of these properties; a system just is “LoT-like” to the extent that it implements them. Both views require methods for determining whether the behavior of a given system reflects properties of LoTs. On the diagnostic view, a further question remains regarding whether observed properties really reflect an underlying LoT.

The authors do not make this distinction and different parts of their argument suggest different interpretations. Here we show that one can't have both, because the consequences of these interpretations are incompatible. For one thing, the choice determines whether neural networks are competitors to the LoTH or candidate implementations.

The authors' central argument seems to place them in the diagnostic camp. They suggest that their six core properties comprise a homeostatic property cluster (Boyd, 1991, 1999). As opposed to prototype concepts, in which features need not be related, a crucial question for homeostatic property clusters is: What maintains the clustering? In the standard case, it is maintained by an underlying homeostatic mechanism (or set of

mechanisms), which causes the properties to co-occur to an unexpected degree (Boyd, 1990). For example, the characteristic properties of a biological species – the paradigm case of a homeostatic property cluster – are diagnostic of that species, and tend to co-occur because of the shared genotype of species members, which is maintained by evolutionary forces. Proposing that the LoT is a homeostatic property cluster evokes an analogy: Some underlying homeostatic mechanism causes the properties of LoTs to cooccur. This mechanism is the extra constitutive component; even if two systems exhibit identical indicators, the presence or absence of the mechanism determines which ones *really are* LoTs. This accords with the authors' treatment of “non-LoT-like architectures such as DNNs” (target article, sect. 3, fn. 5) as a priori incompatible alternatives to LoTs through much of the paper. Although the core properties may emerge in these systems, a difference in (or lack of) the underlying LoT-like mechanism would mean that deep neural networks (DNNs) could never count as LoTs.

If the properties are just diagnostic, what more is needed for an LoT? The authors do not provide a specific proposal for an underlying homeostatic mechanism, but without one it is unclear what rates of co-occurrence of properties the LoTH predicts. In lieu of a specific prediction, the authors suggest that properties should at least co-occur more frequently than one would expect “from a theory-neutral point of view” (target article, sect. 2, para. 13). But what would one expect? The baseline cannot be a “chance” rate of co-occurrence, because the properties that the authors specify are not, in principle, independent. *Predicate–argument structure* seems to presuppose both *role-filler independence* and *discrete constituents*, while having *logical operators* should enable *inferential promiscuity*. Co-occurrence of properties can only be evidence for the LoTH if it is co-occurrence *over and above* the rate implied by their mutual dependence. Without an estimate of this baseline, it is unclear whether the evidence the authors review actually provides a compelling “abductive, empirical argument for LoTH” (target article, sect. 2, para. 13). On a diagnostic view, both what the diagnostic properties are and their expected rate of co-occurrence should ultimately be causally determined by the underlying homeostatic mechanism. The challenge, then, is to characterize that mechanism.

The constitutive view sidesteps this explanatory challenge. On this view, all there is to being LoT-like is exhibiting the relevant properties. There is no further prediction about above-baseline cooccurrence and no need to posit any underlying mechanisms that maintain homeostatic unity. Nor are the LoTH and DNNs incompatible explanatory paradigms competing to account for the same experimental data. Rather, the LoTH highlights important, multiply realizable properties that stand as targets for any representational format to instantiate, and which might emerge in neural networks that are not explicitly augmented with other, more LoT-like mechanisms. This amounts to a form of *compatibilism* about DNNs and the LoTH. At one point, the authors explicitly endorse this position. They write, “neural-network architectures might be able to implement an LoT architecture... Our six core LoT properties help specify a cluster of features that such an implementation should aim for” (target article, sect. 4.3, para. 6), and they are unwilling to suggest any “in-principle limitations of DNNs” (target article, sect. 4.3, para. 6). However, this clashes with their central argument. On a constitutive view, there is no unifying homeostatic mechanism, so the homeostatic cluster collapses into a prototype concept. Moreover,

*Equal contribution.

the pieces of evidence presented by the authors that particular current DNNs fail to manifest LoT properties or explain some phenomenon (e.g., abstract object representations) cannot weigh in favor of the LoTH and against DNNs as cognitive models in principle. They just suggest that those DNNs do not implement LoTs. This is a critical choice point for proponents of the LoTH: You can embrace compatibilism and the constitutive view or endorse the more robust commitments of the diagnostic property cluster account. You cannot have both.

An unresolved question may influence the choice between these two options. Will neural networks need to be augmented with rule-like operations to account for human competences, as the authors suggest? If so, this would favor the diagnostic view, with LoTs as underlying mechanisms that play a strong explanatory role in cognitive architecture. If, on the contrary, DNNs turn out to be able to explain human cognitive capacities without being augmented with other kinds of architectures (as in neuro-symbolic hybrids), that would support the constitutive view and a weaker, guiding role for the LoTH. Far from abandoning it, this result would allow the LoTH to provide its traditional explanatory benefits without requiring implementation in a rule-based system.

Acknowledgments. We are grateful for helpful discussions with Joshua Schechter, Richard Kimberley Heck, Stavros Orfeas Zormpalas, and Randall O'Reilly.

Financial support. This work was supported in part by an SSNAP Fellowship (award no. 383-001202) to S. M. and J. R., funded by the John Templeton Foundation, and by a Jacobs Foundation Fellowship, awarded to R. F.



Competing interest. None.

References

- Boyd, R. (1990). Realism, approximate truth, and philosophical method. In C. W. Savage (Ed.), *Scientific theories* (pp. 355–391). University of Minnesota Press.
- Boyd, R. (1991). Realism, anti-foundationalism, and the enthusiasm for natural kinds. *Philosophical Studies*, 61(1–2), 127–148.
- Boyd, R. (1999). Kinds, complexity, and multiple realization: Comments on Millikan's "Historical kinds and the special sciences." *Philosophical Studies*, 95(1–2), 67–98.

Neither neural networks nor the language-of-thought alone make a complete game

Iris Oved^a, Nikhil Krishnaswamy^b 

James Pustejovsky^c  and Joshua K. Hartshorne^d 

^aIndependent Scholar, 911 Central Ave; San Francisco, CA, USA; ^bDepartment of Computer Science, Colorado State University, Fort Collins, CO, USA;

^cDepartment of Computer Science, Brandeis University, Waltham, MA, USA and

^dDepartment of Psychology and Neuroscience, Boston College, Chestnut Hill, MA, USA

irisoved@gmail.com, irisoved@paradoxlab.org

nikhil.krishnaswamy@colostate.edu, <https://www.nikhilkrishnaswamy.com/>

jamesp@cs.brandeis.edu, <https://jamespusto.com/>

joshua.hartshorne@bc.edu, <http://l3atbc.org/index.html>

doi:10.1017/S0140525X23001954, e285

Abstract

Cognitive science has evolved since early disputes between radical empiricism and radical nativism. The authors are reacting to the revival of radical empiricism spurred by recent successes in deep neural network (NN) models. We agree that language-like mental representations (language-of-thoughts [LoTs]) are part of the best game in town, but they cannot be understood independent of the other players.

Quilty-Dunn et al. have done a service in summarizing major lines of empirical data supporting a role for symbolic, language-like representations (a language-of-thought [LoT], construed broadly) in theories of cognition. This overview is particularly pressing for audiences of the overly hyped popular press on deep neural networks (NNs). However, Quilty-Dunn et al. have done a disservice to LoT by setting unfavorable terms for the debate. In particular, they (1) overlook the fact that an LoT is necessarily part of a larger system and thus its effects should rarely be cleanly observed, and (2) do not address well-known concerns about LoTs.

Quilty-Dunn et al. note that statements in an LoT are formed of discrete constituents and denote functions from possible worlds to truth values. Consider a situation in which Alice beats Bart at tug-of-war. This might be represented in an LoT as BEAT (ALICE, BART, TUG-OF-WAR). Fodor (1975) argued that constituents (BEAT, ALICE, BART, TUG-OF-WAR) are “atomic” (unstructured) pointers to metaphysically real entities: Events (beating), properties (Aliceness, Bartness), kinds (tug-of-war), and so on.

Unfortunately, such entities do not appear to exist; nature is not so easily carved at its joints. An alternative – implicit in many Bayesian models – is to treat the symbols as reifications of some distribution in the world: There are some features that are reliably (if probabilistically) encountered in combination, and we use, for example, “Alice” to refer to one such combination (or a posited essence that explains the combination; see Oved, 2015). This straightforwardly allows for recognition, for example through an NN classifier (Pustejovsky & Krishnaswamy, 2022; Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015). Thus, the LoT sentence BEAT(ALICE, BART, TUG-OF-WAR) means that in observing the referred-to scene, we would recognize (our NN classifier would identify) an Alice, a Bart, a beating, and tug-of-war, and that these entities would be arranged in the appropriate way (see also Pollock & Oved, 2005). (For readers familiar with possible worlds semantics, the proposition picks out the set of possible worlds where all those recognitions would happen.)

This approach explains, for instance, why we tie ourselves in knots trying to decide whether a cat with the brain of a skunk is a cat or a skunk, or whether the first chicken egg preceded or followed the first chicken. In the LoT, SKUNK, CAT, and CHICKEN are reified abstractions tied to recognition procedures. The world is messier, and the recognition procedures sometimes gum up. Note further that different methods for identifying skunks and cats, and so on (NNs, prototypes, inverse graphics, etc.) have characteristic imprecisions if not outright hallucinations. The predictions of any LoT theory cannot be separated from the manner in which the symbols map onto the world.

Reasoning presents additional complications. Most people infer from *Alice beat Bart at tug-of-war* that Alice is stronger, that both are humans not platypodes, are not quadriplegic, and played tug-of-war in a gym or field not while flying. Although none of these inferences necessarily hold, keeping a completely

open mind about them requires willful obtuseness. Critically, such graded, probabilistic inferences have been the bane of symbolic reasoning theories, including LoTs. A promising avenue is to treat LoT statements as conditions on probable worlds generated from a generative model of the world (Goodman, Tenenbaum, & Gerstenberg, 2014; Hartshorne, Jennings, Gerstenberg, & Tenenbaum, 2019). That is, one considers all possible worlds in which Alice beat Bart at tug-of-war. Because the prior probability of aerial quadriplegic tadpoles playing tug-of-war is low, we discount those possibilities (barring additional evidence).

Because we cannot do a census of possible worlds, this process requires an internal model of the world. Thus, the exact inferences one gets depend on not just the LoT but on what one believes about the world. They also depend on the nature of the model. In some domains, symbolic generative models seem to capture human intuitions, whereas in others we seem to use analog simulations (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). For example, when imagining Alice beating Bart at tug-of-war, we might use abstract causal beliefs about tug-of-war (Hartshorne et al., 2019), or we might simulate Alice pulling the rope and Bart dragging along the ground in her direction; the latter is more sensitive to physical properties of the players and the field. Moreover, as a practical matter, one must marginalize out (“average over”) irrelevant parts of one’s world model (e.g., who Bart’s parents are and what he plans to eat after the match). Determining what is relevant is tricky and substantially affects inferences. Indeed, Bass, Smith, Bonawitz, and Ullman (2021) show that some “cognitive illusions” may be explained by biases in how relevance is determined.

Note that if the above approach is right, the categorical behavior often taken as emblematic of LoTs is likely to be masked by the probabilistic, graded natures of the grounding procedure and the model of the world.

So far, we’ve followed the Fodorian atomic treatment of constituents, but this is controversial. Linguists note that words tend to have many distinct meanings: One can throw a book (the physical object) or like a book (usually the content conveyed by the book, not the physical object). One can beat Bart or the bell, but in fundamentally different ways. There are many reasons not to treat these different meanings as homophones (a single word that refers to many unrelated concepts), one of the most obvious being that you end up needing an enormous (potentially unbounded) conceptual library. Perhaps we do, but linguists have noted that there are systematic correspondences between the various meanings, and that this can only be explained if the symbols Fodor takes to be atomic in fact have structure that contributes to meaning and governs their resulting conceptual combination and composition (Jackendoff, 1990; Pustejovsky, 1995). These solutions can be debated, but the problems have to be solved somehow.

Quilty-Dunn et al. provide a useful description of LoTs. Testing LoT theories, however, requires looking beyond the LoT to how it is used within a larger cognitive system. This, in almost all cases, will involve complex trade-offs and interactions with graded, distributed, and analog systems of representation and processing.

Acknowledgments. We thank members of the BabyBAW team, Mengguo Jing, and Wei Li for valuable discussion.

Financial support. Funding was provided by NSF 2033938 and NSF 2238912 to J. K. H., NSF 2033932 to J. P., and ARO W911NF-23-1-0031 to N. K.

Competing interest. None.

References

- Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, 38(7–8), 413–424.
- Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard University Press.
- Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2014). *Concepts in a probabilistic language of thought*. Center for Brains, Minds and Machines (CBMM).
- Hartshorne, J. K., Jennings, M. V., Gerstenberg, T., & Tenenbaum, J. (2019). When circumstances change, update your pronouns. *Cognitive Science* (p. 3472).
- Jackendoff, R. S. (1990). *Semantic structures*. MIT Press.
- Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naive utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, 20(8), 589–604.
- Oved, I. (2015). Hypothesis formation and testing in the acquisition of representationally simple concepts. *Philosophical Studies* 172(1), 227–247.
- Pollock, J., & Oved, I. (2005). Vision, knowledge, and the mystery link. *Philosophical Perspectives*, 19, 309–351.
- Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.
- Pustejovsky, J., & Krishnaswamy, N. (2022). Multimodal semantics for affordances and actions. In Human-Computer Interaction. Theoretical Approaches and Design Methods: Thematic Area, Held as Part of the 24th HCI International Conference, Proceedings, HCII 2022, Virtual Event, June 26–July 1, 2022, Part I (pp. 137–160). Cham: Springer International Publishing.
- Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, 21(9), 649–665.
- Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. *Advances in Neural Information Processing Systems*, 28.

Evidence for LoTH: Slim pickings

David Pereplyotchik 

Department of Philosophy, Kent State University, Kent, OH, USA

dpereply@kent.edu

<https://kentstate.academia.edu/DavidPereplyotchik>

doi:10.1017/S0140525X23002066, e286

Abstract

In this commentary, I contend that a representative sample of the arguments in the target article miss the mark. In particular, the interface problem provides no warrant for positing similarities between representational formats, and the evidence from neurocognitive, animal, and behavioral studies is inconclusive at best. Finally, I raise doubts about whether the authors’ central hypothesis is falsifiable.

In this commentary, my aim is to develop four separate objections to arguments in the target article. These concern (i) interfaces between representational formats, (ii) how to interpret the P600 ERP signature, (iii) the relation between deep neural network (DNN) models and innateness, and (iv) the significance of performance measures in evaluating DNNs.

Let’s begin with what the authors call the “interface problem.” They argue that “if cognition is largely LoT-like, and perception feeds information to cognition, then we should expect at least some elements of perception to be LoT-like, because the two systems need to interface” (target article, sect. 4, para. 2). This claim, though common, is puzzling. If DNN models have demonstrated anything, it’s that virtually any representational format can be transformed into any other, given suitable training. Names can be mapped to faces, spatial arrays to numerical quantities, letters

into phonemes, intentions into motoric instructions, and so on. Moreover, DNNs routinely do this in ways that appear not to be sensitive to any syntactic properties of the interfacing representations (Arbib, 2003). This strongly suggests that two interfacing systems need not have much, if anything, in common with one another, regardless of whether either of them is language-of-thought (LoT)-like.

If that's correct, then it raises the larger question of what the interface problem was ever supposed to be. The issue has been heavily studied in the theory of action, but, tellingly, the prominent solutions in this area often *multiply* the number of representational formats, introducing a new kind of demonstrative concept (Butterfill & Sinigaglia, 2014) or a “motor schema” that mediates between intentions and low-level motoric instructions (Mylopoulos & Pacherie, 2017). As Christensen (2021) in effect points out, the question of *how* such representations are mapped into one another is not going to be answered by reference to their representational format(s). The substantive questions are how such mappings arise and what happens on the occasions when they fail. Plausible answers to these questions will likely appeal to learning, innate endowment, and low-level neurocognitive mechanisms, but *not* to format.

Turn now to some of the neurocognitive evidence that the authors marshal. They argue that “structured relations in scene grammar display curious hallmarks of language-like formats. For instance, the P600 ERP increases for syntactic violations in language, and also increases for stimuli that violate visual scene ‘syntax’” (target article, sect. 4.2.2, para. 4). However, the P600 has a variety of interpretations, and not all these fit neatly with the authors' reasoning. For instance, the P600 may be a trigger for *conscious* reevaluation of a stimulus that has caused processing issues, regardless of whether the underlying representational system is LoT-like (Batterink & Neville, 2013; van Gaal et al., 2014). More importantly, the P600 has been shown to reflect incongruence or discordance in tonal music (Featherstone, Morrison, Waterman, & MacGregor, 2013), demonstrating that a representational format – in this case, that of musical cognition – need not involve predication, logical operations, or automatic inferential promiscuity in order to induce a P600 response.

The authors might reply that musical cognition is demonstratively sensitive to recursive structure and that it exhibits filler-role relations (Lerdahl & Jackendoff, 1983). The representational format involved is, thus, arguably LoT-like. But this response raises the deeper question of what it would take to falsify their main proposal. If musical cognition meets only *half* of the criteria that they take to be indicative of an LoT-like format, does this constitute a refutation of their hypothesis that such criteria naturally cluster together? If not, then what would?

Let's now consider issues in animal cognition. The authors argue that the paucity of relevant input to a newborn chick's visual system prior to an experiment “points away from DNN-based explanations of abstract object representations” (target article, sect. 5.1, para. 5). The idea seems to be that, if a representational capacity is innate, rather than acquired through some type of learning, then DNN models of this capacity are superfluous. But this argument runs together two separate issues – LoT versus DNN, on the one hand, and learning versus innateness, on the other. Although proponents of a DNN modeling do tend to lean empiricist, this sociological fact can be misleading. In actuality, fans of DNN-style representational formats need to have no commitment whatsoever on the issue of innateness. It

could well be that a chick, or any other critter, inherits a “frozen” pretrained DNN-style representational system as a part of its genetic endowment. Presumably, in the real world, such a system would have been “trained” into its innate structure over the course of the creature's evolutionary past – a process akin to selecting a particularly successful DNN out of several and then using it as a “seed” for training a new cohort of variants.

Before closing, let me draw attention to the authors' use of performance data in evaluating deep convolutional neural network (DCNN) models. On the one hand, they argue that “divergence between DCNN and human performance echoes independent evidence that DCNNs fail to encode human-like transformation-invariant object representations” (target article, sect. 5.1, para. 6). On the other, they are steadfastly committed to a competence/performance distinction, which renders the evidence that they cite questionable. As Firestone (2020) points out, performance measures are often unreliable guides in assessing the psychological plausibility of a DCNN, whether in vision or in any other domain. In psycholinguistics, performance has long ceased to be a reliable sign of human competence (Pereplyotchik, 2017), and computational linguists disagree about what performance measures to use (e.g., Sellam et al., 2022), even in DCNNs that make no claim to psychological plausibility. Thus, in order to make their case for the inadequacy of DCNN models – again, in vision or any other domain – the authors would need to cite evidence that evaluates the *competence* of such models. How to do this is, at present, far from a settled matter, so the performance measures they rely on are almost certain to be equivocal.

In summary, a representative sample of the arguments in the target article simply fail. The interface problem provides no warrant for positing similarities between representational formats, and the evidence from neurocognitive, animal, and behavioral studies is inconclusive at best. It is, moreover, unclear whether the authors' central hypothesis is falsifiable.


Acknowledgments. My thanks to Jacob Berger, Daniel Harris, and Myrto Mylopoulos for helpful discussion.

Competing interest. None.

References

- Arbib, M. A. (Ed.). (2003). *The handbook of brain theory and neural networks*. MIT Press.
- Batterink, L., & Neville, H. J. (2013). The human brain processes syntax in the absence of conscious awareness. *Journal of Neuroscience*, 33(19), 8528–8533.
- Butterfill, S., & Sinigaglia, C. (2014). Intention and motor representation in purposive action. *Philosophy and Phenomenological Research*, 88(1), 119–145.
- Christensen, W. (2021). The skill of translating thought into action. *Review of Philosophy and Psychology*, 12, 547–573.
- Featherstone, C. R., Morrison, C. M., Waterman, M. G., & MacGregor, L. J. (2013). Semantics, syntax or neither? A case for resolution in the interpretation of N500 and P600 responses to harmonic incongruities. *PLoS ONE*, 8(11), 1–13.
- Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences of the United States of America*, 117(43), 26562–26571.
- Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. MIT Press.
- Mylopoulos, M., & Pacherie, E. (2017). Intentions and motor representations: The interface challenge. *Review of Philosophy and Psychology*, 8(2), 317–336.
- Pereplyotchik, D. (2017). *Psychosyntax: The nature of grammar and its place in the mind*. Springer.
- Sellam, T., Yadlowsky, S., Wei, J., Saphra, N., D'Amour, A., Linzen, T., ... Pavlick, E. (2022). The MultiBERTs: BERT reproductions for robustness analysis. *ICLR*. [arXiv]. arXiv: [arXiv:2106.16163](https://arxiv.org/abs/2106.16163).
- van Gaal, S., Naccache, L., Meuwese, J. D. I., van Loon, A. A. M., Leighton, A. H., Cohen, L., & Dehaene, S. (2014). Can the meaning of multiple words be integrated unconsciously? *Philosophical Transactions of the Royal Society*, 369, 20130212.

Using the sender–receiver framework to understand the evolution of languages-of-thought

Ronald J. Planer 

School of Liberal Arts, Faculty of Arts, the Social Science and Humanities,
University of Wollongong, Wollongong, NSW, Australia

rplaner@uow.edu.au

https://scholars.uow.edu.au/display/ronald_planer

doi:10.1017/S0140525X23002078, e287

Abstract

This commentary seeks to supplement the case Quilty-Dunn et al. make for the psychological reality of languages-of-thought (LoTs) in two ways. First, it focuses on the reduced physical demands which LoT architectures often make compared to alternative architectures. Second, it embeds LoT research within a broader framework that can be leveraged to understand the evolution of LoTs.

Quilty-Dunn et al. adduce evidence for the psychological reality of languages-of-thought (LoTs) from a wide range of empirical domains. Their case inherits support from each domain, while depending on none. This is a powerful way to make such a case. Their article is, moreover, timely. It is a most welcome antidote to the steady rise in antirepresentationalist sentiment in many philosophy of cognitive science circles in recent years. Overarching theories of cognition that eschew any role for computational procedures applied to structured symbols are not serious contenders unless and until they adequately account for detailed empirical information of the sort discussed by Quilty-Dunn et al.

So, my impression of their article is strongly positive. Here, my aim is to supplement their case in two ways. First, by drawing attention to a distinct empirical rationale for LoTs. And second, by situating LoT research within a broader framework that promises to shed light on the evolution of LoTs.

LoT-based architectures often make much reduced physical demands compared to alternative architectures. Symbols are constructed in a combinatorial fashion, and their sequence properties play a role in individuating symbols. This allows for efficient representation. Additionally, the meaning of a (complex) symbol is a function of that symbol's parts, together with their mode of composition. Symbols, to some extent, analytically deconstruct their referents. Such symbols allow for the use of compact computational procedures (as opposed to, say, lookup tables). Together, these principles can reduce the demand on physical resources (e.g., neurons) by orders of magnitude.

These points have been most forcefully argued by Gallistel and colleagues (Gallistel, 1990, 2008; Gallistel & King, 2011), often with examples drawn from animal cognition. A good case is the caching behavior of western scrub jays. These birds are estimated to encode the location of thousands of caches (Clayton & Krebs, 1995). Moreover, for each location, they encode what was cached, when it was cached, and whether they were watched while caching it (their caches are often pilfered) (Clayton, Yu, & Dickinson, 2001). Additionally, they make flexible use of this information (e.g., to retrieve cached items in an efficient way) (Clayton

et al., 2001). Arguably, scrub jays could not physically realize the requisite symbols and computations except by instantiating an LoT. And even if they could, an LoT architecture might still have been selectively favored for its increased economy. Brain tissue is expensive, after all.

But how might such symbol systems evolve in the first place? Progress on this question can be made by using the “sender–receiver framework.” This framework is inspired by the signaling games first presented by David Lewis (1969). At their simplest, a signaling game features a sender who can observe the variable state of the world and send a signal (but cannot act), and a receiver who can observe the signal (but not the world), and act. Acts have consequences for both sender and receiver, and both have preferences regarding which act should be done when. Lewis showed that, given certain conditions (e.g., rationality, common interest, common knowledge), informative signaling can arise and stabilize. Decades later, these games were revisited by Skyrms who showed how Lewis's constraints could be significantly relaxed (Skyrms, 1995, 2004, 2010). Indeed, Skyrms showed how even completely *mindless* agents can evolve informative signaling under many conditions.

Skyrms's generalization of the Lewis model allows us to apply that model *within* organisms, not just *between* them (Godfrey-Smith, 2014; Planer, 2019; Planer & Godfrey-Smith, 2021). Two cognitive mechanisms (or one and the same cognitive mechanism at different times) can serve as sender and receiver in a Lewis–Skyrms-style setup. And this allows us to see (with the aid of the theory and results that have grown up around signaling games in recent decades) how signaling systems, including rather complex ones, can arise and stabilize in brains over phylogenetic and ontogenetic timeframes. This includes systems that are plausibly conceived of as LoTs (Planer, 2019).

Using the sender–receiver framework, Planer and Godfrey-Smith (2021) present a taxonomy of signs displaying different forms of structure (Table 1). Unfortunately, there is not scope here to go through the details of this taxonomy. Suffice it to say that the taxonomy is structured by two tripartite distinctions among signs, namely, *atomic-composite-combinatorial*, and *nominal-organized-encoding*, which are envisaged as plausible, incremental evolutionary pathways. On this taxonomy, an LoT is a sign system (used in cognition) that is simultaneously combinatorial and encoding. As a *combinatorial* sign system, it contains signs that are constructed out of other signs belonging to the system (and hence, there is sharing of parts across signs), and moreover, the order of the parts of a sign matters to how the sign functions in communication and/or computation. And as an *encoding* sign system, there is a systematic principle (or set of such principles) that assigns meaning to complex signs based not only on the identity of their parts, but also on where those parts occur in the sign (and so, particular locations within a complex sign have meaning). It is combinatoriality that allows for maximally efficient representation and encoding principles that allow for the use of compact, efficient algorithms. These properties are very close to those Quilty-Dunn et al. call “discrete constituency” and “role-filler independence” (while “predicate–argument structure” [target article, sect. 2, para. 9] can be understood as a special case of encoding).

A final methodological point. The sender–receiver framework is closely associated with a family of formal signaling models. And although the orientation to sign use that the framework fosters is not inherently formal (Planer & Godfrey-Smith, 2021), these models are very useful. For they make testing ideas about the

Table 1 (Planer). Taxonomy of signs based on their formal and semantic structure

	Nominal	Organized	Encoded
Atomic	No sharing of parts across signs; no use of relations among signs. <i>Examples:</i> Most animal alarm call systems.	No sharing of parts across signs; some use of relations among signs. <i>Examples:</i> An alarm call system where call intensity relates to distance of predator in a rough way.	No sharing of parts across signs; sign form governed by an encoding principle. <i>Examples:</i> An alarm call system where call intensity precisely maps distance of predator.
Composite	Sharing of parts across signs; order of parts does not matter; no role for relations among signs. <i>Examples:</i> A bird song system in which receivers attend only to the presence or absence of species-typical syllables.	Sharing of parts across signs; order of parts does not matter; some use of relations among signs. <i>Examples:</i> Early stages of a honeybee waggle dance; a bird song in which receivers attend only to overall song complexity.	Sharing of parts across signs; order of parts does not matter; sign form governed by an encoding principle. <i>Examples:</i> Actual honeybee waggle dance.
Combinatorial	Sharing of parts across signs; order of parts matters; no role for relations among signs. <i>Examples:</i> Some great-ape gesture sequences (e.g., an attention getter + point to genitals); “pyow-hacks” in putty-nosed monkeys.	Sharing of parts across signs; order of parts matters; some use of relations among signs. <i>Examples:</i> Human languages words (English nouns share more parts with other nouns than verbs, and vice-versa); DNA triplet code (optimized for error minimization).	Sharing of parts across signs; order of parts matters; governed by an encoding principle (that gives a semantic role to sign parts). <i>Examples:</i> Human language sentences, sentences in a language of thought; genes.

Adapted from Planer and Godfrey-Smith (2021).

emergence of various forms of structure tractable. Research on the evolution of LoTs can no doubt benefit from these formal tools. Most obviously, signaling models might be used to investigate whether and under what conditions Quilty-Dunn et al.’s six core properties indeed cluster (or form subclusters). Additionally, such models might be used to test the idea that LoTs evolve at interfaces between other systems, as interface systems can be naturally modeled as intermediaries in so-called signaling chains.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

Clayton, N. S., & Krebs, J. R. (1995). Memory in food-storing birds: From behaviour to brain. *Current Opinion in Neurobiology*, 5(2), 149–154.
 Clayton, N. S., Yu, K. S., & Dickinson, A. (2001). Scrub jays (*Aphelocoma coerulescens*) form integrated memories of the multiple features of caching episodes. *Journal of Experimental Psychology: Animal Behavior Processes*, 27(1), 17.
 Gallistel, C. R. (1990). *The organization of learning* (Vol. 336). MIT Press.
 Gallistel, C. R. (2008). Learning and representation. *Learning and Memory: A Comprehensive Reference*, 1, 227–242.
 Gallistel, C. R., & King, A. P. (2011). *Memory and the computational brain: Why cognitive science will transform neuroscience*. John Wiley.

- Godfrey-Smith, P. (2014). Sender–receiver systems within and between organisms. *Philosophy of Science*, 81(5), 866–878.
- Lewis, D. (1969). *Convention: A philosophical study*. Harvard University Press.
- Planer, R. J. (2019). The evolution of languages of thought. *Biology & Philosophy*, 34, 1–27.
- Planer, R. J., & Godfrey-Smith, P. (2021). Communication and representation understood as sender–receiver coordination. *Mind & Language*, 36(5), 750–770.
- Skyrms, B. (1995). *Evolution of the social contract*. Cambridge University Press.
- Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge University Press.
- Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford University Press.

Language-of-thought hypothesis: Wrong, but sometimes useful?

Adina L. Roskies^a  and Colin Allen^b 

^aDepartment of Philosophy, Dartmouth College, Hanover, NH, USA and
^bDepartment of History & Philosophy of Science, University of Pittsburgh,
 Pittsburgh, PA, USA
adina.roskies@dartmouth.edu, colin.allen@pitt.edu
<https://faculty-directory.dartmouth.edu/adina-l-roskies>, <https://www.hps.pitt.edu/people/colin-allen>

doi:10.1017/S0140525X23001991, e288

Abstract

Quilty-Dunn et al. maintain that language-of-thought hypothesis (LoTH) is the best game in town. We counter that LoTH is merely one source of models – always wrong, sometimes useful. Their reasons for liking LoTH are compatible with the view that LoTH provides a sometimes pragmatically useful level of abstraction over processes and mechanisms that fail to fully live up to LoT requirements.

Quilty-Dunn et al. ask the question “What is the format of thought?” (target article, sect. 1, para. 1). Their answer is that it is language-like. Despite the title of their paper, and their contention in the introduction that “LoTH is the *best* game in town” (italics in original; target article, sect. 1, para. 6), Quilty-Dunn et al. don’t explicitly argue for this superlative conclusion except with respect to one study of human concept learning, discussed in section 3. Rather, they show how to interpret various studies of human and animal cognition as involving representational formats that have (most of) the sixfold homeostatic property cluster they use to characterize language-of-thought (LoT). We concede that models using LoT-formatted representations may be very useful, not just for human concept learning but also for the capacities of nonhuman animals and some artificial systems. But the claim that language-of-thought hypothesis (LoTH) is the best game in town requires a more substantial defense than Quilty-Dunn et al. provide, including more specificity about the hypothesis itself and an answer to the question “Best for what?”

We should all grant that some human cognitive processes operate over and with natural language representations. After all, we internalize the culturally acquired structures of language, logic, and mathematics, and regiment our thinking accordingly. We can explicitly think thoughts in linguistic form, so some processes must represent and operate on language-like structures, including discrete constituents, logical operations, predicate–argument structure, and so on. Certainly, models that involve

language-like representations can reproduce psychological data; this is especially so when the tasks are posed in language and engage rules of inference that exploit predicate–argument structure, and so on, in execution. Because LoTH is itself modeled on characteristics of natural language, these processes will a fortiori be well modeled by LoTH. However, Quilty-Dunn et al. make a more sweeping claim: They argue that the utility of LoT is pervasive and the best way to model thought even in nonlinguistic creatures. Here, too, we note that if you describe your tasks and results in language-like ways, you may find models that incorporate language-like elements natural to turn to. We contend that the utility of LoTH for abstractly characterized cognitive processes does not entail that LoTH is the best way to model these processes in more detail, or even at the same level of abstraction, unless what counts as an LoT is so attenuated as to be nearly universally applicable.

To the extent that their examples suggest that LoTH is the best game in town, it is because their homeostatic property cluster view of LoT allows them to embrace most representational formats, including many never before conceived of as examples of the LoT. For instance, they argue that object files possess at least five of the six features of LoT representations, and they use this to support their claim that the LoT format is a better fit to object files than Carey’s iconic account. But if this is so, then the same seems to hold for the feature maps in Triesman’s (1998) attentional model, and LoTH seems to have been weakened almost to the point of vacuity. In Triesman’s model, a spotlight of attention potentiates processing in cognate spatial regions of disparate feature maps. This model is not particularly language-like: It is not generative, not serial, not recursive, and does not have discrete word-like tokens. However, the feature binding can be seen as implementing predicate–argument structure (“these features collocated in this spatial location”), the maps represent discrete features with abstract content like color and shape that can be reused in different bindings, providing a kind of role-filler independence that allows the system to bind the same features to different objects and different features to the same object. However, if map-like models are also LoT models, then the hypothesis fails to differentiate between cognitively very different kinds of solutions – it is too general to do much work. And if map-like models are not instances of the LoT, then Quilty-Dunn et al. have failed to argue that their LoT-based account of object files is superior to one based on icons or maps.

The above discussion highlights the need for clearer statements about if and when the six properties characteristic of LoT are instantiated. As the example of object files shows, their approach is to give an LoT-inspired interpretive dance after the theory is already on offer. However, terms such as “discrete” and “inferentially promiscuous” (target article, sect. 3, para. 6) are too vague to guide cognitive scientists in constructing theories that make testable predictions about behavior and neural mechanisms in humans and other animals. This does not entail that they are useless for picking out a class of models, but class membership then is largely in the eye of the beholder. Moreover, Quilty-Dunn et al. miss two important points from the philosophy-of-science of modeling. One is that the formal expression of a model has features that should not be attributed to the target system (Andrews, 2021; Beer, [in prep.](#)). The other is that models serve particular scientists’ purposes more or less well. We think it salutary to recall Box’s maxim that “all models are wrong, but some are useful” (Box, 1976). All models are approximations to the phenomena they model, and different models highlight different aspects of those phenomena. For example, Bayesian models involve computations over probabilities assigned to propositions

or sets of propositions, in terms of priors on hypotheses and statements of evidence. Those models explicitly calculate posteriors using discrete elements and logical operators. Quilty-Dunn et al. cite these as another instance of the successes of LoTH. However, few, if any, think that our brains explicitly represent Bayes theorem or calculate the posteriors by crunching numbers (except when forced to in math classes, etc.). Rather, Bayes theorem is thought to be an analytical optimal representation that is only approximated by brain mechanisms that work according to different principles. If one is interested in how the brain computes Bayes-like posteriors, LoT may not be helpful at all. Thus although LoT may be a useful model for some phenomena for some applications, it will not be for others. It is certainly not the only game in town, nor is it always the best.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

- Andrews, M. (2021). The math is not the territory: Navigating the free energy principle. *Biology & Philosophy*, 36(3), 30. doi:10.1007/s10539-021-09807-0
- Beer, R. B. (in prep.). On the proper treatment of dynamics in Cognitive Science.
- Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799. doi:10.1080/01621459.1976.10480949
- Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 1373, 1295–1306. doi:10.1098/rstb.1998.0284

Linguistic meanings in mind

Alexis Wellwood^a  and Tim Hunter^b

^aSchool of Philosophy, University of Southern California, Los Angeles, CA, USA and ^bDepartment of Linguistics, University of California, Los Angeles, Los Angeles, CA, USA

wellwood@usc.edu, <https://semantics.land>

timhunter@ucla.edu, <https://timhunter.humspace.ucla.edu/>

doi:10.1017/S0140525X23001887, e289

Abstract

The target article focuses on evidence from nonlinguistic faculties to defend the claim that cognition generally traffics in language-of-thought (LoT)-type representations. This focus creates needed space to discuss the mounting accumulation of nonclassical evidence for LoT, but it also misses relevant work in linguistics that directly offers a perspective on specific hypotheses about candidate LoT representations.

Quilty-Dunn et al. defend the claim that, roughly, mentation generally traffics in language-of-thought (LoT)-type structures. Classically, such evidence has been drawn from considerations of human language understanding. If that were the best (or indeed only) evidence for the claim, though, it would be relatively easy for an LoT skeptic to dismiss it. Quilty-Dunn et al. therefore focus on evidence from nonlinguistic domains, where signs of language-like representations are all the more striking. Although the ease of connecting the dots between natural language and LoT-type structures makes the study of linguistic cognition a

poor choice for arguing that minds generally implement *some* sort of LoT-type representations, we submit that it also makes human language a particularly fruitful domain for formulating and defending *specific* hypotheses about candidate LoT representations. And indeed, Quilty-Dunn et al. say little about this beyond the useful identification of general properties of LoTs; nor do they say specifically how “logicality” should be understood in the context of potentially significantly modular minds. In this note, we introduce into this discussion a recent strand of theoretical and experimental work on (broadly) logical expressions in linguistic semantics that directly bears on these matters.

First, let us suppose along with Quilty-Dunn et al. that believing something consists, at least in part, of entertaining a particular LoT representation (Field, 1978). Then, a natural way to explain the finding that, for example, “[t]elling participants... they will see a pairing of a group with pictures of pleasant (or unpleasant) things is much more effective at fixing implicit attitudes than repeatedly pairing the group and the pleasant/unpleasant things” (target article, sect. 6.2, para. 5), is to suppose that the outputs of (specifically linguistic) language comprehension simply are LoT expressions (e.g., Hunter & Wellwood, 2023; Wellwood, 2020). On the contrary, the pathway from associationistic learning episodes to such representations is rather less direct. Indeed, this view on adult linguistic understanding pairs well with views about language acquisition that presuppose human beings come equipped with a shared conceptual system; on such views, learning the meanings of words involves solving a mapping problem rather than a concept acquisition problem (Gleitman, 1990).

Second, if beliefs in the relevant sense are syntactically structured objects, then the relevant theories should say something about their structure. We understand this to be a question in the spirit of the distinction between the computational- versus algorithmic-level (Marr, 1982). Quilty-Dunn et al. cite work implementing probabilistic languages-of-thought (PLoTs) which is taken to “provide[] defeasible evidence that some approximation of the computational elements of the model are realized in human cognitive architecture,” but, as they note, “further evidence is needed to establish” its “algorithmic-level reality” (target article, sect. 3, note 4).

Recent research in “psychosemantics” explicitly aims to approach these lower levels of abstraction – at least down to “Level 1.5” (Peacocke, 1986). It begins by noting the possibility of specifying boundlessly many truth-conditionally equivalent, but intensionally distinct (Church, 1941) characterizations of the meaning of sentences like *Most of the Cs are Bs* and *There are more As than Bs* with the putatively logical items *most* or *more*, and asks which best correspond to how people represent them. Specifically, given a formal characterization φ of the meaning of sentence *S*, we can probe whether people are biased to make use of the information explicitly called for in φ . For example, the two expressions “ $|C \& B| > |C \setminus B|$ ” and “ $|C \& B| > |C| - |C \& B|$ ” equally well capture the truth-conditions of *Most of the Cs are Bs*, but the former calls for the cardinality of $C \setminus B$ (i.e., the *Cs* which are not *Bs*) while the latter calls for the cardinality of *C*. Carefully manipulating which perceptual-cognitive information is readily available in a task where subjects must evaluate the truth of *S*, one can check the resulting fit between what people draw on when evaluating *S* and what φ calls for.

The most striking findings of this research are that English speakers are indeed remarkably uniform in the information they recruit to evaluate a sentence like *Most As are Bs*, and they are pretty stubborn in those preferences even when other strategies are readily available (Hackl, 2009), more accurate (Pietroski, Lidz, Hunter, & Halberda, 2009), or more extensible (Lidz,

Pietroski, Halberda, & Hunter, 2011). Furthermore, these findings replicate for speakers evaluating translational equivalents in Polish (Tomaszewicz, 2011), and they are systematically different from those recruited for sentences with *more* under extensionally equivalent circumstances (e.g., Knowlton et al., 2021).

Finally, Quilty-Dunn et al. cite evidence for logical reasoning as evidence for LoT, but we are not told what it means for a mind to “use logical operators” (target article, sect. 5.2, para. 1) which are “generalizab[le] across content domains” (target article, sect. 1, para. 7). For the latter, it must be that LoT (sub-)expressions can interface with distinct and potentially modular systems, but this raises the question of what ties together the various domain-specific interpretations of a single LoT expression. To concretize the problem, suppose that the end result of understanding *There are more apples than bananas* is the LoT expression “ $A > B$ ” and that of *There is more sand than mud* is “ $S > M$.” Specifically, what’s required is a specification of (i) how the single symbol “ $>$ ” (Wellwood, 2019) can be interpreted by cognitive systems operating both over domains representing pluralities of objects and those representing stuff (Odic, Pietroski, Hunter, Lidz, & Halberda, 2012; see Rips & Hespos, 2015), and (ii) the logical relationships that an LoT expression enters into by virtue of being built around this symbol in a certain way (e.g., via proof-theoretic inference rules). We have suggested (Hunter & Wellwood, 2023) that for a nonlinguistic system to interpret a symbol such as “ $>$ ” in the appropriate way is exactly for this interpretation to abide by some algebraic laws that are, in effect, specified by some inference rules governing expressions built out of “ $>$.” For example, a rule licensing a logical inference from “ $A > B$ ” and “ $B > C$ ” to “ $A > C$ ” essentially specifies that, in each content domain where “ $>$ ” has an interpretation, that interpretation must correspond to a transitive binary relation.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

- Church, A. (1941). *The calculi of lambda conversion*. Princeton University Press.
- Field, H. (1978). Mental representation. *Erkenntnis*, 13(1), 9–61.
- Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1(1), 3–55.
- Hackl, M. (2009). On the grammar and processing of proportional quantifiers: *Most* versus *more than half*. *Natural Language Semantics*, 17, 63–98.
- Hunter, T., & Wellwood, A. (2023). *Linguistic meanings interpreted*. Forthcoming in the Proceedings of the 45th Annual Conference of the Cognitive Science Society.
- Knowlton, T., Hunter, T., Odic, D., Wellwood, A., Halberda, J., Pietroski, P., & Lidz, J. (2021). Linguistic meanings as cognitive instructions. *Annals of the New York Academy of Sciences*, 1500, 134–144.
- Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of *most*. *Natural Language Semantics*, 6(3), 227–256.
- Marr, D. (1982). *Vision*. MIT Press.
- Odic, D., Pietroski, P., Hunter, T., Lidz, J., & Halberda, J. (2012). Young children’s understanding of “more” and discrimination of number and surface area. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(2), 451–461.
- Peacocke, C. (1986). Explanation in computational psychology: Language, perception and level 1. *Mind and Language*, 1(2), 101–123.
- Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of “most”: Semantics, numerosity and psychology. *Mind and Language*, 24(5), 554–585.
- Rips, L., & Hespos, S. J. (2015). Divisions of the physical world: Concepts of objects and substances. *Psychological Bulletin*, 141(4), 786–811.
- Tomaszewicz, B. (2011). Verification strategies for two majority quantifiers in Polish. In I. Reich, E. Horch, & D. Pauly (Eds.), *Proceedings of Sinn und Bedeutung* (Vol. 15, pp. 597–612). Universaar – Saarland University Press.
- Wellwood, A. (2019). What more is. *Philosophical Perspectives, Philosophy of Language*, 32(1), 454–486.
- Wellwood, A. (2020). Interpreting degree semantics. *Frontiers in Psychology*, 10, 1–14.

Toward biologically plausible artificial vision

Mason Westfall 

Department of Philosophy, Philosophy–Neuroscience–Psychology Program,
Washington University in St. Louis, St. Louis, MO, USA
w.mason@wustl.edu
<http://www.masonwestfall.com>

doi:10.1017/S0140525X23001930, e290

Abstract

Quilty-Dunn et al. argue that deep convolutional neural networks (DCNNs) optimized for image classification exemplify structural disanalogies to human vision. A different kind of artificial vision – found in reinforcement-learning agents navigating artificial three-dimensional environments – can be expected to be more human-like. Recent work suggests that language-like representations substantially improves these agents’ performance, lending some indirect support to the language-of-thought hypothesis (LoTH).

Image classifiers implemented with deep convolutional neural networks (DCNNs) have been taken by many to tell against language-of-thought (LoT) architectures. Quilty-Dunn et al. argue that this is a mistake. These image classifiers exhibit deep structural disanalogies to human vision, so, whether or not they implement LoT architectures tells us little about human vision. This is perhaps unsurprising, because biological vision is plausibly not optimized solely for image classification (Bowers et al., 2022, p. 10). Would training artificial vision under more ecologically realistic conditions produce a more realistic model of human vision? To make progress on this question, I describe some reinforcement-learning (RL) agents trained to navigate artificial three-dimensional environments on the basis of how things appear from their perspective, and explain why we might expect their vision to be more human-like. Interestingly, language-like representations seem to be especially helpful to these agents. They explore more effectively, more quickly learn novel tasks, and are even facilitated in downstream image classification. These models arguably provide some indirect evidence for the language-of-thought hypothesis (LoTH) about human vision, and may offer some clues as to why LoT architectures arose evolutionarily.

What is biological vision optimized for, and what would artificial vision that was similarly optimized be like? One answer to the first question is that biological vision is optimized for an agent’s success in their environment. Success requires a number of competences that vision must contribute to simultaneously. Agents need to effectively explore, learn new behaviors, and act to achieve their goals, all while the environment changes in often surprising ways.

Recent work in RL arguably more closely approximates the optimization problem facing biological agents. Artificial RL agents can learn to do many complex tasks, across a variety of environments – most interestingly, in this context, exploring and pursuing goals in artificial three-dimensional environments like Habitat (Savva et al., 2019), Matterport3D (Chang et al., 2017), Gibson Env (Xia et al., 2018), Franka Kitchen (Gupta et al., 2019), VizDoom (Kempka et al., 2016), Playroom (Tam

et al., 2022), and City (Tam et al., 2022). One way of accomplishing this – especially in environments where environmental reward is sparse – is by making novelty intrinsically rewarding. These “curious agents” can learn, without supervision, representations that enable them to perform navigation tasks, interact with objects, and also perform better than baseline in image recognition tasks (Du, Gan, & Isola, 2021). As the authors put it, their agents are “learning a task-agnostic representation for different downstream interactive tasks” (Du et al., 2021, p. 10409).

One challenge these researchers face is how to characterize novelty. Superficial differences in viewing angle or pixel distribution can easily be rated as highly novel, leading to low-level exploration that does not serve learning conducive to achieving goals. A recent innovation is to equip RL agents with “prior knowledge, in the form of abstractions derived from large vision-language models” (Tam et al., 2022, p. 2). Doing so enables the state space over which novelty is defined to be characterized by abstract, semantic categories, such that novelty is defined in task-relevant ways (Mu et al., 2022). This method has been shown to substantially improve performance across a variety of tasks and environments, compared to nonlinguistic ways of characterizing the state space (Mu et al., 2022; Schwartz et al., 2019; Tam et al., 2022). The improvements are especially pronounced for tasks involving relations between objects, for example, “Put an OBJECT on a {bed, tray}” (Tam et al., 2022, p. 2), reminiscent of work on relations reviewed in the target article (Hafri & Firestone, 2021). As the authors note, their training on vision–language representations that encode “objects and relationship” instead of on ImageNet – optimized for classification – should be expected to be more successful (Tam et al., 2022, p. 10).

Why would linguistic categories facilitate performance? One possibility is that language compresses the state space in ways that facilitate successful actions. The semantic categories enshrined in natural language tend to abstract from action-irrelevant variation, and respect action-relevant variation. So, visual processing optimized relative to natural language categories is *de facto* optimized for action-relevant distinctions. The LoT architecture characteristic of object files and visual working memory seems well-suited to serving this function (though LoT plausibly is importantly different from natural languages; Green, 2020; Mandelbaum et al., 2022). Predicating abstract properties of individual objects in a LoT is poised to guide action, because abstract semantic categories often determine the action affordances available for some individual object, independent of nuisance variation associated with, for example, viewing angle (though viewing angle is plausibly relevant for more fine-grained control tasks; Parisi et al., 2022, p. 6). Such abstract, task-agnostic representations are also able to transfer to new tasks or environments, in which familiar kinds take on novel relevance for action.

These recent innovations in RL arguably offer indirect support for the LoTH as applied to humans. Of course, similar performance can be achieved by distinct underlying competence, and we should not exaggerate how similar even artificial RL agents’ performance actually is to humans at present. Nevertheless, language-like structures appear especially helpful for artificial agents when faced with rather more biologically plausible optimization problems than the one that faces image classifiers. Perhaps an LoT served our ancestors similarly in an evolutionary context. Language-like structures enabled creatures to encode abstract properties in a task-agnostic way, which nevertheless facilitated downstream performance on a wide variety of tasks, as the environment changed. It’s not hard to imagine why evolution might set to it that such a system stuck around.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.

References

- Bowers, J. S., Malhorta, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... Blything, R. (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 1–74.
- Chang, A., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., ... Zhang, Y. (2017). Matterport3D: Learning from RGB-D data in indoor environments. arXiv preprint, arXiv:1709.06158.
- Du, Y., Gan, C., & Isola, P. (2021). *Curious representation learning for embodied intelligence*. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10408–10417.
- Green, E. J. (2020). The perception–cognition border: A case for architectural division. *Philosophical Review*, 129(3), 323–393.
- Gupta, A., Kumar, V., Lynch, C., Levine, S., & Hausman, K. (2019). Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. arXiv preprint, arXiv:1910.11956.
- Hafri, A., & Firestone, C. (2021). The perception of relations. *Trends in Cognitive Sciences*, 25(6), 475–492.
- Kempka, M., Wydmuch, M., Runc, G., Toczek, J., & Jaśkowski, W. (2016). *Vizdoom: A doom-based AI research platform for visual reinforcement learning*. 2016 IEEE Conference on Computational Intelligence and Games (CIG). IEEE, pp. 1–8.
- Mandelbaum, E., Dunham, Y., Feiman, R., Firestone, C., Green, E., Harris, D., ... Quilty-Dunn, J. (2022). Problems and mysteries of the many languages of thought. *Cognitive Science*, 46(12), e13225.
- Mu, J., Zhong, V., Raileanu, R., Jiang, M., Goodman, N., Rocktäschel, T., & Grefenstette, E. (2022). Improving intrinsic exploration with language abstractions. arXiv preprint, arXiv:2202.08938.
- Parisi, S., Rajeswaran, A., Purushwalkam, S., & Gupta, A. (2022). *The unsurprising effectiveness of pre-trained vision models for control*. International Conference on Machine Learning. PMLR, pp. 17359–17371.
- Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., ... Batra, D. (2019). *Habitat: A platform for embodied AI research*. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9339–9347.
- Schwartz, E., Tennenholtz, G., Tessler, Chen, & Mannor, S. (2019). Language is power: Representing states using natural language in reinforcement learning. arXiv preprint, arXiv:1910.02789.
- Tam, A. C., Rabinowitz, N. C., Lampinen, A. K., Roy, N. A., Chan, S. C. Y., Strouse, D., ... Hill, F. (2022). Semantic exploration from language abstractions and pretrained representations. arXiv preprint, arXiv:2204.05080.
- Xia, F., Zamier, A., He, Z., Sax, A., Malik, J., & Savarese, S. (2018). *Gibson Env: Real-world perception for embodied agents*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), 2018, pp. 9068–9079.

Is core knowledge in the format of LOT?

Fei Xu 

Department of Psychology, University of California, Berkeley, CA, USA

fei_xu@berkeley.edu

<https://www.babylab.berkeley.edu/feixu>

doi:10.1017/S0140525X23001966, e291

Abstract

Object individuation provides a test case for the claim that infants already have a prelinguistic language-of-thought (LOT). By 12 months, infants represent several sortal-kinds: Object, agent, animate, and perhaps artifact. Infants have also encountered many words for object kinds, animals, people, and artifacts, therefore it remains a viable hypothesis that language learning may play a causal role in the acquisition of sortal-kinds, contra Quilty-Dunn et al.

Quilty-Dunn et al. put forth a strong thesis: That language-of-thought hypothesis (LOTH) is the “best game” in town as far as a computational theory-of-mind is concerned. They marshal evidence from object perception, deductive reasoning, and other domains to support this claim. I am sympathetic to the view that LOT continues to provide a philosophical and conceptual foundation for modern cognitive science. In this commentary, however, I submit, as I did in Xu (2019), that core knowledge systems in human infants do not satisfy the criteria for being in the format of LOT. Here I focus on the domain of object, in particular object individuation in human infants.

Inspired by an analysis of the logic of common nouns (Macnamara, 1986; Wiggins, 1980), we reported a series of experiments demonstrating that 10-month-old infants failed to use sortal-kind distinctions (e.g., between a duck and a ball) to establish a representation of two objects in an occlusion event; by 12 months, they can do so (the “is-it-one-or-two” task; Xu & Carey, 1996; see Xu, 1997, 2007, for reviews). We argued that it is not until the end of the first year that infants represent basic-level sortal-kinds such as *duck*, *ball*, *spoon*, and *cup*, and learning a natural language – specifically words for these sortal-kinds – may play a causal role in acquiring these concepts. A lot has happened since then.

For the rest of our discussion, it is important to keep in mind that three pieces of evidence are needed to claim that infants represent sortal-kind concepts in an LOT format: (1) success in using between-kind distinctions in object individuation, (2) failure in using within-kind distinctions in object individuation at the same age, and (3) evidence showing that infants detect the perceptual distinctions between sequentially presented objects over occlusion. In support of the claim that infants younger than 12 months do not represent basic-level sortal-kinds, Xu and Carey (1996) and Xu, Carey, and Quint (2004) presented evidence for (1)–(3). Since then, many published studies have used similar methods (the “is-it-one-or-two” task) to further investigate the ontogenetic origin of sortal-kind concepts, focusing on three other superordinate-level concepts: *Agent*, *animate*, and *artifact* (see Croteau, Cheries, & Xu, forthcoming, for a review). For the concept of an agent, Bonatti, Frot, Zangl, and Mehler (2002) found that 10-month-old infants successfully individuated a doll head from an inanimate object (a between-kind distinction, agent vs. object), and a doll head from a dog head, but they failed to individuate a doll head from another doll head (a within-kind distinction). Recent studies by Bródy, Oláh, Király, and Biro (2022), Taborda-Osorio, Lyons, and Cheries (2019), and Taborda-Osorio and Cheries (2018) found that 10-, 11-, or 13-month-old infants used preferences, social-moral dispositions, and internal properties to individuate agents. For the concept of animacy, Surian and Caldi (2010) found that 10-month-old infants successfully individuated an animate and an inanimate object (a dynamic caterpillar and a stationary cup; a between-kind distinction) but failed to individuate two animates (a rabbit and a bee; a within-kind distinction). Decarli, Franchin, Piazza, and Surian (2020) provided converging evidence, further disentangling the use of sortal-kind versus featural information. Lastly, Futó, Téglás, Csibra, and Gergely (2010) found that 10-month-old infants successfully individuated an object with a function and another object with a different function, although they did not demonstrate a difference in individuation contrasting between-kind versus within-kind distinctions. The studies on agent, animacy, and artifact did not present direct evidence that infants encoded the various relevant

perceptual feature differences, but given what we know about infant perception in general, most would agree that not encoding perceptual differences between objects was an unlikely explanation for the failures in individuation tasks (though see Kibbe & Leslie, 2019). It is also important to note that Wilcox, Baillargeon, Lin, Stavans, and their colleagues have conducted many related experiments over the years, with a strong focus on investigating when infants use *featural* information in object individuation and the relationship between object files and physical reasoning. Their studies have not aimed to probe the development of sortal-kind concepts (e.g., Lin et al., 2021; Stavans, Lin, Wu, & Baillargeon, 2019; Wilcox & Baillargeon, 1998). A review of their studies and the various methodological differences between their methods and the “is-it-one-or-two” task is beyond the scope of this commentary; however, these details are important for interpreting this body of research.

The studies reviewed above support the view that toward the end of the first year of life, infants represent sortal-kind concepts: *Object*, *agent*, *animate*, and perhaps *artifact*. During the first year of life, infants also hear many, many words that refer to basic-level object kinds, people, animals, and artifact kinds. Given the evidence on how words facilitate object categorization, individuation, and inductive inference of nonobvious properties (see Perszyk & Waxman, 2018; Xu, 2002, 2007; and others for reviews), it remains a viable hypothesis that it is language learning that changes the format of early representations into an LOT.

The core knowledge view (Spelke, 2022) also argues for several other systems of early knowledge besides object. In particular, the number sense presents another strong case that these prelinguistic representations are incompatible with an LOT format. A rich body of research suggests that prelinguistic representations of number share very little with the conceptual representations needed for learning number words. It is an open question whether the prelinguistic representations of agents, places, or social beings are in the format of an LOT.

I applaud Quilty-Dunn et al. for drawing our attention, once again, to the significance of the LOTH. If core knowledge systems are indeed not in the format of an LOT, as I have argued here, cognitive scientists face a major challenge in understanding learning and development in many domains: How does language learning change the format of prelinguistic representations, or alternatively, how does language learning create new conceptual representations that are in the format of an LOT?

Acknowledgments. I thank Jenna Croteau and Erik Cheries for helpful discussion.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. None.



References

- Bonatti, L., Frot, E., Zangl, R., & Mehler, J. (2002). The human first hypothesis: Identification of conspecifics and individuation of objects in the young infant. *Cognitive Psychology*, 44(4), 388–426. <https://doi.org/10.1006/COGP.2002.0779>
- Bródy, G., Oláh, K., Király, I., & Biro, S. (2022). Individuation of agents based on psychological properties in 10 month-old infants. *Infancy*, 27(4), 809–820. <https://doi.org/10.1111/INFA.12472>
- Croteau, J., Cheries, E., & Xu, F. (forthcoming). The development of kind concepts: Insight from object individuation and beyond. *Annual Review of Developmental Psychology*.
- Decarli, G., Franchin, L., Piazza, M., & Surian, L. (2020). Infants’ use of motion cues in object individuation processes. *Journal of Experimental Child Psychology*, 197, 104868. <https://doi.org/10.1016/J.JECP.2020.104868>

- Futó, J., Téglás, E., Csibra, G., & Gergely, G. (2010). Communicative function demonstration induces kind-based artifact representation in preverbal infants. *Cognition*, 117(1), 1–8. <https://doi.org/10.1016/j.cognition.2010.06.003>
- Kibbe, M. M., & Leslie, A. M. (2019). Conceptually rich, perceptually sparse: Object representations in 6-month-old infants' working memory. *Psychological Science*, 30, 362–375.
- Lin, Y., Li, J., Gertner, Y., Ng, W., Fisher, C. L., & Baillargeon, R. (2021). How do the objectfile and physical-reasoning systems interact? Evidence from priming effects with object arrays or novel labels. *Cognitive Psychology*, 125, 101–136. <https://doi.org/10.1016/j.cogpsych.2020.101368>
- Macnamara, J. (1986). *A border dispute*. MIT Press.
- Perszyk, D. R., & Waxman, S. R. (2018). Linking language and cognition in infancy. *Annual Review of Psychology*, 69, 231–250.
- Spelke, E. S. (2022). *What babies know*. Oxford University Press.
- Stavans, M., Lin, Y., Wu, D., & Baillargeon, R. (2019). Catastrophic individuation failures in infancy: A new model and predictions. *Psychological Review*, 126(2), 196–225. <https://doi.org/10.1037/REV0000136>
- Surian, L., & Caldi, S. (2010). Infants' individuation of agents and inert objects. *Developmental Science*, 13(1), 143–150. <https://doi.org/10.1111/j.1467-7687.2009.00873>
- Taborda-Osorio, H., & Cheries, E. W. (2018). Infants' agent individuation: It's what's on the insides that counts. *Cognition*, 175, 11–19. <https://doi.org/10.1016/j.COGNITION.2018.01.016>
- Taborda-Osorio, H., Lyons, A. B., & Cheries, E. W. (2019). Examining infants' individuation of others by sociomoral disposition. *Frontiers in Psychology*, 10, 1271. <https://doi.org/10.3389/FPSYG.2019.01271/BIBTEX>
- Wiggins, D. (1980). *Sameness and substance*. Harvard University Press.
- Wilcox, T., & Baillargeon, R. (1998). Object individuation in infancy: The use of featural information in reasoning about occlusion events. *Cognitive Psychology*, 37, 97–155. <https://doi.org/10.1006/COGP.1998.0690>
- Xu, F. (1997). From Lot's wife to a pillar of salt: Evidence for *physical object* as a sortal concept. *Mind and Language*, 12, 365–392.
- Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, 85, 223–250. [https://doi.org/10.1016/S0010-0277\(02\)00109-9](https://doi.org/10.1016/S0010-0277(02)00109-9)
- Xu, F. (2007). Sortal concepts, object individuation, and language. *Trends in Cognitive Sciences*, 11, 400–406. <https://doi.org/10.1016/J.TICS.2007.08.002>
- Xu, F. (2019). Towards a rational constructivist theory of cognitive development. *Psychological Review*, 126, 841–864. <http://dx.doi.org/10.1037/rev0000153>
- Xu, F., & Carey, S. (1996). Infants' metaphysics: The case of numerical identity. *Cognitive Psychology*, 30, 111–153. <https://doi.org/10.1006/COGP.1996.0005>
- Xu, F., Carey, S., & Quint, N. (2004). The emergence of kind-based object individuation in infancy. *Cognitive Psychology*, 49, 155–190. <https://doi.org/10.1016/J.COGPSYCH.2004.01.001>

Authors' Response

The language-of-thought hypothesis as a working hypothesis in cognitive science

Jake Quilty-Dunn^a , Nicolas Porot^b
and Eric Mandelbaum^c 

^aDepartment of Philosophy and Philosophy-Neuroscience-Psychology Program, Washington University in St. Louis, St. Louis, MO, USA; ^bAfrica Institute for Research in Economics and Social Sciences, Mohammed VI Polytechnic University, Ben Guerir, Morocco and ^cDepartment of Philosophy and Department of Psychology, The Graduate Center & Baruch College, CUNY, New York, NY, USA

quiltydunn@gmail.com; sites.google.com/site/jakequiltydunn/
nicolasporot@gmail.com; nicolasporot.com
eric.mandelbaum@gmail.com; ericmandelbaum.com

doi:10.1017/S0140525X23002431, e292

All authors contributed equally; authorship is in reverse alphabetical order.

Abstract

The target article attempted to draw connections between broad swaths of evidence by noticing a common thread: Abstract, symbolic, compositional codes, that is, language-of-thoughts (LoTs). Commentators raised concerns about the evidence and offered fascinating extensions to areas we overlooked. Here we respond and highlight the many specific empirical questions to be answered in the next decade and beyond.

We are extremely grateful for the commentaries we have received which further language-of-thought hypothesis (LoTH) in ways we couldn't delve into in the target article. Some of our commentators criticize the specifics of our proposal; others criticize the wisdom of the endeavor altogether; still others extend the theory in novel and creative ways. We are moved by the time, effort, and thought our 30 commentators put into these commentaries. A proper response to these challenges and extensions would require a book-length format. Here we try to address some of the deep, important, and thorny issues our interlocutors brought up.

The commentaries clustered around a few main topics: Development, object representations in perception, natural language, and the theoretical foundations of LoTH. We begin, fittingly, with development.

R1. LoT and development

The topic of development arose frequently, raised by **Canudas-Grabolosa, Martín-Salguero, & Bonatti (Canudas-Grabolosa et al.); Carey; Cesana-Arlotti; Colombo; Demetriou; Hochmann; Kibbe; Planer; and Xu**. Many theories ask difficult, important questions about the specifics of developmental trajectory. For example Carey, Hochmann, and to some extent Canudas-Grabolosa et al. all have varying levels of skepticism about whether certain logical concepts (e.g., OR) or modal concepts (e.g., POSSIBLE) are available for preverbal infants. Before addressing the specifics, we want to stress how healthy this debate is. Regardless of where one comes down on any specific proposal, these criticisms highlight a serious difference between our LoT and Fodor's. For Fodor, there was no role for developmental psychology; LoT was fixed innately, and the substantive empirical questions were largely restricted to facts about the timescale of developmental triggering. By contrast, we make no claims about radical nativism (or triggering vs. availability at birth, for that matter). Although we are inclined to share the Harvard nativist view (e.g., the views of Spelke, Carey, and Xu) of core cognition, we need not agree with Fodor that all lexeme-sized LoT representations are unlearned.

As **Kibbe** points out, our LoTH framework allows us to ask fine-grained questions about the developing child's shifting representational repertoire. We needn't just ask "does the steady state of adult conceptual mastery exceed the infant's expressive power?" Instead we can also ask which concepts are available pre-linguistically, and which ones are acquired through a process of **Carey**-style bootstrapping; we can discuss stage developments without taking on a Piagetian framework, but while allowing that the innate conceptual endowment needn't be fixed; we can in principle accommodate a full rational constructivist framework (Xu, 2019). Moreover, because some core LoT properties are gradable (both individually and as subclusters), some might emerge before others, or more quickly, and with nonmonotonic shifts

in expressive power. Our framework therefore allows for distinct LoTs not only across species and systems, but also across different stages of development. As **Canudas-Grabolosa et al.** write, LoT is “probably a genus, but, we would add, one whose actual species are still barely known.” We agree. The project of uncovering specific LoTs and their expressive powers at various stages of development and evolution is a challenging one. Perhaps negation and disjunction are available prelinguistically, or perhaps they must be acquired through word learning.

This is an exciting time for the study of logical reasoning in infants and animals, and the commentaries on this topic demonstrate how rapidly the field has evolved – including in the time since we submitted our target article. Although we agree with **Carey** that the debate on baby and animal logic is still open, we are optimistic about the explanatory prospects of LoTH. Carey proposes baboons and children between 17 months and 3 years old sequentially simulate possibilities in three- and four-cup tasks (Leahy, Huemer, Steele, Alderete, & Carey, 2022). This would predict their roughly 50–50 responding, where mental-logic accounts appeal to imperfect performance. However, near-ceiling performance on such tasks can be found in other species (Pepperberg, Gray, Cornero, Mody, & Carey, 2019), as can the holding of multiple alternatives in mind at once on a task similar to the two-cup task (Engelmann et al., 2021). The latter result is a hurdle for sequential simulation, and the former suggests that genuine disjunction is indeed possible without language.

Carey also notes two-cup task failures for 14- and 15-month olds (Feiman, Mody, & Carey, 2022) and suggests that success starts at the same time as negation acquisition in some languages: 17 months. But if success has something to do with natural-language acquisition, it is curious that we should find a 50–50 pattern of results in three-cup tasks for children as old as three, well after even English learners have started producing linguistic negation. Why not near-ceiling performance? There are limits to logical reasoning even here (performance constraints or a preference for suboptimal strategies), and they could be masking competence in younger toddlers and infants. Relatedly, why might negation acquisition induce simulation? The 50–50 responding and negation production both emerge at around 17 months, but if there is an inference to be drawn here, it is that linguistic negation should improve logical reasoning, not that it primes sequential simulation. Why acquiring linguistic negation would improve one’s powers of simulation is unclear.

One might wonder whether 50–50 responding is because of a response bias to pick a container from either side half the time. However, Leahy et al. (2022) report that 3-year olds who are asked to “throw away” one of the containers in a three-cup task picked from the pair 81% of the time. There might be other response biases at play here, however. Humans like symmetry and dislike lopsidedness, and a 3-year old might prefer not to throw away the lone container, instead throwing away one of the containers in the pair, thus leaving one container on each side. One way to substantiate this explanation would be to do a pure throw-away task: For example, with one candy on the left and two candies on the right, does a 3-year old pick a candy to throw away randomly, or is there a preference to throw away one from the pair? If so, that would suggest a response bias can explain the Leahy et al. results.

Recent evidence suggests that easing performance constraints can eliminate 50–50 responding in 3-year olds in otherwise similar tasks. Alderete and Xu (2023) developed a task involving transparent gumball machines, one of which *might* produce the

desired gumball and the other of which *must*. This task places much lighter demands on working memory, and does not require the arguably demanding feat of quantifying over trajectories of balls through Y-shaped tubes as in other recent studies. Alderete and Xu found that 3-year olds select the correct gumball machine roughly 90% of the time. We think these results tentatively support optimism about performance-error-based explanations of earlier findings regarding disjunction (specifically, problems with working-memory-demanding tasks, including tracking hidden locations and anticipating trajectories, and response biases).

Other tasks (Cesana-Arlotti et al., 2018; Cesana-Arlotti, Kovács, & Téglás, 2020) yield higher success rates, even in 12-month olds. As **Cesana-Arlotti** notes in his commentary, these results, and in particular the pupil dilation results, are difficult to accommodate with simulation. **Carey** instead appeals to 1–1 mapping of object files to percepts. Because adults show the same pupillometric profile as 12-month olds in such studies, her interpretation suggests adults, like infants, resolve this problem with 1–1 mapping. This is an area for future research that can help us distinguish between 1–1 mapping and logical explanations of these results. Because adults can perform disjunctive syllogism (DS) in Cesana-Arlotti et al.’s task, the 1–1 mapping explanation opens up the intriguing possibility that multiple redundant reasoning processes are carried out by adults. Independent evidence regarding the cognitive mechanism behind the pupillary dilation and eye movements can help shed light on whether 12-month olds are indeed performing DS.

Even in adults, as **Lupyan** points out, “rule-based reasoning is far more difficult than it should be if such logical operators actually underlie much of our perception and reasoning.” This is an important point that allows us to clarify our view. We agree, of course, that people make systematic errors in reasoning. We have seen ourselves how difficult it can be to teach symbolic logic to undergraduates. But we think the cognitive architecture of belief, rather than its format, explains many deviations from norms of reasoning (Mandelbaum, 2019). Unsurprisingly, the architecture we favor can operate over representations in an LoT (Porot & Mandelbaum, 2020, 2022; Quilty-Dunn & Mandelbaum, 2018). We do not assume that logical operators of LoTs are just like those of formal or even natural languages; in fact, we agree with **Canudas-Grabolosa et al.**, **Cesana-Arlotti**, and **Wellwood & Hunter**, who argue they differ (see also Mandelbaum et al., 2022; Porot, 2019).

Xu pushes back on our defense of LoT-like effects in the use of object files for physical reasoning in infancy, including research grounded in her landmark work on this topic (especially Xu & Carey, 1996). Xu’s commentary makes valuable points about the methodological issues, especially concerning differences in the evidence for abstract representation of a special class of superordinate kinds (e.g., *object* vs. *agent*; for related philosophical work, see Murez & Smortchkova, 2014; Westfall, forthcoming) and ordinary basic-level categories like *knife* and *marker*. We are grateful for these points.

We grant that the evidence is not decisive about abstract representation of basic-level categories before 12 months, and that careful attention to experimental details is required to make progress on this issue. We add two points here. The first concerns the relevance of the Stavans and Baillargeon (2018), Stavans, Lin, Wu, and Baillargeon (2019), and Lin et al. (2021) results. Children’s failures to use basic-level categories for object individuation before 12 months might be because of a performance error. Appeals to

performance errors are unhelpful without specifying the relevant performance constraints and the experimental paradigms that might overcome them. We suggest that the aforementioned results implicate “catastrophic individuation failures” (Lin et al., 2022), wherein featural *and* categorical information encoded in the object file system fail to be used in physical reasoning, resulting in disagreement between the two systems about the number of occluded objects and thus no coherent expectation on the infant’s part. This account predicts that priming relevant information – including surface features like color or the function of a basic-level artifact category like *knife* or *marker* – helps the object file system to make the relevant information accessible to physical reasoning. There is considerable independent evidence from vision science that the object file system makes some encoded properties available for use in object individuation and not others, and that drawing attention to a feature increases the likelihood that it is made available (see Quilty-Dunn & Green, 2023). We agree that further work along the lines Xu describes is needed to show decisively that appropriate priming can allow infants to use abstract basic-level categories for object individuation.

Our second point concerns the Nc event-related potential (ERP) used by Pomiechowska and Gliga (2021). In their experiment 2, 12-month-old infants were unfamiliar with labels for a group of basic-level categories (e.g., *feather*, *guitar*), as confirmed by their parents and by their insensitivity to the category in experiment 1. Infants in the experimental condition saw multiple instances of two categories sorted by category without a linguistic category label, and those in the control condition were shown the category instances without category-relevant training. They found that “[i]nfants who learned nonverbal categories prior to the EEG task displayed sensitivity to across-category but not to within-category object changes” (Pomiechowska & Gliga, 2021, p. 9). This result offers hope that infants represent abstract basic-level categorical information in object files. Questions for future research include: How early can these ERP results be observed? What, if any, role do the representations in this experiment play in the Xu and Carey’s (1996) “is-it-one-or-two” task? Would similar nonverbal category training enable success on that task before 12 months? These questions are experimentally tractable, and we are excited to see the trajectory of developmental research on the format of object representations in the years to come.

R2. Object representations

Many commentators focused on our discussion of object file representations, which was just one component of our discussion of LoT in perception. Some directly rejected our claims (Block; Lupyán; Xu), some worried that the argument overextends (Attah & Machery; Cheng; Roskies & Allen), and some saw opportunity for testable claims about development (Carey; Kibbe; Xu), which we discuss in section 1 of the target article. We are thankful to have such a rich interdisciplinary discussion on the representational format of this core posit of contemporary cognitive science in these commentaries.

Block worries that the object file representations that can be maintained in visual working memory (VWM) are distinct representations with distinct formats from object representations used in online vision. He cites evidence that perception relies on iconic representations of objects, while VWM involves a distinct LoT-like form of object representation. There is a background dispute here regarding the border between perception and cognition, which Block (2023) argues is because of format. Two of us

(Mandelbaum, 2018; Quilty-Dunn, 2020c) argue that format cannot explain the perception–cognition border because there are non-iconic, conceptual, LoT representations in vision. For our purposes, we could concede the “border” issue and limit ourselves to the claim that *visual cognition* involves LoT-like object files. In fact, however, we think the evidence suggests that the same object files that can be held in VWM are genuinely visual – they underwrite clearly visual phenomena like apparent motion (Odic, Roth, & Flombaum, 2012), multiple-object tracking (Haladjian & Pylyshyn, 2008), and much more (see Green, 2023, for an overview and reply to Block’s apparent motion case). It is possible that some “object-like” phenomena in vision (e.g., some gestalt phenomena, figure-ground segregation) might involve iconic representations of regions of space and their interaction with attention guided by full-blown object files. We are grateful to Block for years of discussion on the structure of perception and look forward to continuing this conversation, hopefully with more and more relevant experimental evidence.

Cheng argues that the case for LoT structures in tactile perception is “at least as good as the object file” case, but that this “generates a potential worry” that our notion of LoTH is too weak. However, we don’t agree that Cheng’s commentary provides convincing evidence for any of our six LoT properties in tactile perception, nor that this evidence is as strong as the case we provide for LoT structures in object perception. We appreciate the opportunity to demonstrate the rather demanding constraints on positing LoT structures that our six core properties entail.

His argument for discrete constituents in tactile representations is that touch involves “multiple tactile stimuli, each of them exists independent of one another.” This is not evidence for discrete constituents, however. To support the presence of discrete constituents, there would need to be evidence (i) that representations of these stimuli are *composed* into a complex structure of which they are *constituents*, and (ii) that the representations remain *discrete* while being composed (to rule out, e.g., holistic feature composition, tensor products, and other nondiscrete forms of composition). The mere presence of representations of different stimuli that exist independently does not meet these constraints. Compare: One might have a mental map of Brooklyn and a mental map of St. Louis, which exist independently of each other, but that fact by itself does not entail the presence of discrete constituents (cf. Camp, 2018). Similarly, the fact that tactile stimuli “exhibit different properties at different times” does not demonstrate that they exhibit predicate–argument structure, with a representation of a predicate and a distinct representation of an argument that predicate applies to. And the fact that geometrical properties like lines and triangles are represented in touch does not show that their encoding abstracts away from low-level tactile details, which would be needed to infer abstract conceptual content. We are optimistic that there might be evidence for these LoT properties in tactile perception, but the mere presence of multiple stimuli that can change properties and represent geometrical shapes is insufficient for establishing that fact.

Similarly, Roskies & Allen argue that object files are among the examples “never before conceived of as examples of the LoT” and that our defense of LoT models of object files entails that feature maps as understood in Treisman’s feature integration theory are also LoT structures, and thus that “LoTH seems to have been weakened almost to the point of vacuity.” As a historical point, we note that the claim that researchers in the area have never conceived of the idea that object files might be LoT representations is inaccurate. Carey (2011, p. 116) and Xu (2019) have

conceived of the idea enough to argue explicitly against it, and Pylshyn (2009) and Cavanagh (2021) have argued in favor of it. Furthermore, feature maps in Treisman's theory (which should be distinguished from current-day theories of visual attention, such as Guided Search 6.0; Wolfe, 2021) do not appear to satisfy any of our six LoT properties. Roskies & Allen say that feature maps represent "discrete features" but cite no evidence that the features are represented discretely (rather than, say, via analog magnitude representations; Clarke, 2022); indeed they mention that a feature map "does not have discrete word-like tokens." As pointed out in response to **Cheng** above, the mere presence of two separate representations (e.g., two feature maps) does not provide evidence of discrete constituents. They also assert that feature maps encode "abstract content" but their examples are "color and shape." As we understand the abstract conceptual content property of LoTs, modality-specific formats representing specific colors and shapes, as feature maps do, are paradigmatic cases of representations that *lack* abstract conceptual content.

Roskies & Allen appear to conflate feature maps, which represent individual features and not their conjunctions, with the outputs of the feature integration operation. This distinction is important, because for Treisman the outputs of feature integration are the very object file representations at issue (Kahneman, Treisman, & Gibbs, 1992). For example, Roskies & Allen write that "*the feature binding* can be seen as implementing predicate-argument structure" (emphasis added) and that feature maps provide "a kind of role-filler independence that allows the system to bind the same features to different objects and different features to the same object." The mere presence of a feature binding operation does not require (i) that the output (i.e., the representation of the feature conjunction) represents the features via discrete constituents rather than holistically; (ii) that the features are predicates applied to an explicitly represented argument rather than in an icon that lacks a discrete constituent that stands for an individual property-bearer; or (iii) that the features obey role-filler independence rather than other forms of composition (e.g., the sort instantiated by deep convolutional neural networks [DCNNs] that encode feature conjunctions; Taylor & Xu, 2021). All of these are nontrivial empirical claims. Indeed, some forms of feature binding in the visual system seem to be iconic and possess none of these LoT-like properties (Quilty-Dunn, *forthcoming*).

The distinction between the initial detection of separate features (e.g., feature maps) and the way features are represented in the output of a binding operation (e.g., object files) is also relevant to the critique from **Attah & Machery**. They argue that our six LoT properties are vaguely defined (to a certain extent, we agree; see sect. R5) and that they are "too readily discoverable in cognition." Their example pertains to role-filler independence and binding features to objects: "even the swapping of visual features to objects (e.g., misattributing the color of one object to another) counts as a demonstration of role-filler independence."

This description is not quite right. One might have parallel systems for detecting individual features (as in Treisman & Gelade, 1980) and then a binding operation that constructs representations of objects and their features without role-filler independence (e.g., because conjoined features are represented in a final holistic map without discrete constituents for each feature). This system could generate illusory conjunctions (Treisman & Schmidt, 1982), where a blue square and red circle are misperceived as a blue circle and red square because of errors in the binding operation, without role-filler independence. The evidence we discuss

in our perception section, crucially, concerns how features are represented *after* being encoded in object files; that is, after the compositional operation of binding is completed. Without illusory conjunctions in the initial encoding of feature conjunctions, individual features (including not only color and shape but also more abstract properties like the openness or closedness of a book) are represented discretely enough that they can be swapped from one item to another in VWM. Thus the object file representation must compose various features while allowing that these representations can easily be separated from one another and swapped to different objects. This kind of format is a nontrivial instance of role-filler independence.

This evidence could have turned out differently. Object files could have encoded feature conjunctions without preserving discrete representations of features that incur their own individual memory costs, as was originally thought to be the case in research on object-based VWM storage (Luck & Vogel, 1997). It could have turned out that feature representations, once encoded into object files, tend to degrade together in VWM rather than feature-by-feature, as an LoT model predicts – and so on for the other properties of object file representations detailed in the target article. If the evidence had differed in these ways, then there would be no support for LoT models of object files.

Despite our disagreements, we thank **Cheng, Roskies & Allen**, and **Attah & Machery** for raising these objections. The specific examples of tactile perception, feature maps, and illusory conjunctions allow us to illustrate in detail how our defense of LoT structures in the object file system does not trivially apply to these other cases.

R3. LoTH and natural language

Our target article focuses on sources of evidence for LoTH other than natural language. But natural language is of course deeply linked to the LoTH and the relation is not merely evidential. Many of the commentaries discuss promising avenues of research on LoT and natural language.

Canudas-Grabolosa et al. claim that "rather than regarding [natural languages] as the origin of logical abilities in thought, one could look at their semantics as crystalized repositories of thought primitives." We very much agree, and much could be done to better understand both the process by which this "crystalization" occurs and the precise lessons it offers for the structure of individual LoTs. But as **Cesana-Arlotti** and **Wellwood & Hunter** point out, natural language cannot be taken as a simple blueprint for mental syntax. LoTs of particular cognitive systems and (especially) those of nonhuman LoTs might differ syntactically from any known natural or formal languages. Similarly, perhaps some syntactic features of natural language differ from those of prelinguistic LoT(s), and as **Dupre** suggests, such structures might scaffold human-specific cognitive abilities. **Xu** argues that this scaffolding from natural language might extend to LoT structures in core cognition, including object file representations (though see sect. R1 for our reply). Until we know more about the syntactic features and representational primitives of actual LoTs, we cannot know very much for sure how close or far they are from natural language.

These are the very early beginnings of an expansive research program that seems to open as many questions as it settles. For example, a programmatic approach to understanding the syntactic and representational primitives of thought would allow us to type cognitive systems, stages of development, and even minds

themselves (across species) with remarkable fineness of grain. This in turn would raise new questions about the origins of each of those LoTs. Fishes can have the same fin shape despite sharing no ancestor that has it, because there are only a few good ways to swim when one has a spine. Similarly, perhaps certain syntactic features, such as predicate–argument structure or logical connectives, are the products of convergent evolution, having reemerged repeatedly because of their usefulness to biological organisms faced with similar constraints and problems to solve. Relatedly, if there are very many possible LoTs, how many are *not* worth considering when testing between alternatives? In this regard, principles of natural and formal languages can be a useful (if imperfect) starting point for understanding the fundamental properties of an LoT.

We agree with **Oved, Krishnaswamy, Pustejovsky, & Hartshorne (Oved et al.)** that eventually LoT will need to be understood as a single element among a broader system in which it is embedded. In particular, questions that Fodor thought of as exotica – for example, how LoTs interact with images, analog magnitudes, motor intentions (Mylopoulos, Pacherie, & Shepherd, *MS*; Shepherd, 2019), and various formats useful for reasoning, categorization, and interacting with the world – will have to be reckoned with. We think a major area for research in the coming years will be how different formats interact and coexist, such as dual codes in perception (Quilty-Dunn, 2020c). However, we hesitate to endorse all of Oved et al.’s recommendations. Although we can agree that concepts like CAT are “tied to recognition procedures,” we doubt that constituents of an LoT sentence are “reified abstractions” over these recognition procedures or distributions of worldly features.¹ Instead, we suspect LoT concepts are genuine atoms of thought, not composed of features; we distinguish between concepts and the features that trigger their deployment, as Fodor did (Fodor, 1998). This is an in-house dispute, however; LoTH itself is compatible with the more pragmatist approach adopted by Oved et al. as well as the antipragmatist view we are drawn to.

Oved et al. also raise the problem of polysemy, that is, the flexibility of reference observed in many ordinary words, such as “bottle” in “Mary drank the bottle” and “Mary smashed the bottle” (Pustejovsky, 1995, is a *locus classicus*; see also Vicente, 2018). Fodor dismissed polysemy as either nonexistent or an uninteresting instance of homonymy, as when “bank” can refer to a financial institution or land alongside a river (Fodor & Lepore, 1998). Our defense of LoTH in the target article is officially neutral on this issue, but in fact we take the issue seriously. Unlike homonymy, polysemy allows for flexibility in anaphoric reference (“Parched and belligerent, Mary drank the bottle and then smashed it”) and copredication (“Lunch was delicious and informative”) (Murphy, 2021). Unlike homonymy, forms of so-called “regular” (i.e., systematic and rule-governed) polysemy – like, for example, using a word for a container or its contents as in “bottle,” or for an animal and the meat from that animal as in “chicken,” or for an aperture and the object that fills it as in “window” – show up robustly across languages (Srinivasan & Rabagliati, 2015).

So what do we say about an LoT concept like BOTTLE? One option, as **Oved et al.** note, is to insist that each referent has a distinct LoT concept, and therefore to allow one word in the thinker’s lexicon to address many concepts (Carston, 2012; Pietroski, 2018). Another option is to drop Fodor’s referentialist requirement on LoT symbols, and allow one concept BOTTLE to function as a pointer to a memory location where diverse bodies of

information can be retrieved to resolve polysemy inherent in the concept itself (Quilty-Dunn, 2021). Yet another option posits large nonatomic concepts, pieces of which can be deployed on different occasions (Ortega-Andrés & Vicente, 2019) – though this option perhaps plays least well with LoTH. We take the issue of polysemy and LoT concepts to be largely open and amenable to empirical investigation. As with many questions about the way trains of thought unfold, the answer likely lies in the interaction of many representational formats, organized in nontrivial ways by LoTs.

Much discussion of natural language in the commentaries presupposed that some LoTs might predate it, including in animals. **Kaufmann & Newen** sum up our target article by saying that we propose LoTH can “explain all animal cognition” and that we “suggest understanding all communication and reasoning through language-like structures in a wide sense, to justify compositionality.” They point to orangutan long calls, which can be explained with non-LoT representations. As we say in the target article, however, we did not intend to suggest that LoT is the “only game in town,” nor did we make any claims about communication, in orangutans or any other creature (including humans); indeed we explicitly denied that all cognition in human or nonhuman animals is explicable through LoT-like formats. Instead, we pointed to evidence for LoTs in many corners of the animal kingdom, focusing on specific experimental paradigms and cognitive domains such as cup tasks and physical reasoning. Because Kaufmann & Newen did not discuss these paradigms or domains, we are left wondering which aspects of our application of LoTH to nonhuman animals they find implausible. In any case, there is a rich research program ahead exploring how LoTs and other formats divide the mind’s labor, including in nonhuman animals.

Antony points to applications of LoT to person-level phenomena involving beliefs and other propositional attitudes. We wholeheartedly agree. Although our aim was to focus on explanatory successes of LoTH in areas more remote from explicit thought, we think some of the strongest evidence for LoTH remains its utility in explaining the structure of belief. We also concur that dispositionalism and antirepresentationalism about belief struggle to explain familiar phenomena like opacity (why we can believe that *p* under one description but not another) and the enormous conceptual gap between belief and behavior (how the belief that *p* can cause us to engage in incompatible behaviors depending on the other attitudes we use in inference). In other projects, all three of us have argued for a full-throated representationalism about belief (Mandelbaum, 2014, 2016; Porot & Mandelbaum, 2020; Quilty-Dunn & Mandelbaum, 2018). We are thankful to Antony for bringing these classic issues to the fore, and we are optimistic that LoTH will continue to prove useful in solving problems from the structure of bee cognition up to Frege’s puzzle.

Antony’s comments on belief can be extended to implicit cognition, which we did in section 6 of the target article. We just here note that applying LoT and belief to the study of implicit attitudes has been an enormously fruitful paradigm, as can be seen by the groundbreaking work of Benedek Kurdi and **De Houwer**. We note this as it’s easy to take for granted how quickly the study of implicit attitudes has changed. Ten years ago, associationism for understanding attitudes still reigned. As **Madva** implies, our view has now mostly become the accepted backdrop in the experimental implicit bias literature. The recent history of implicit attitude research thus exemplifies the serious power of LoTH.

R4. What is LoTH committed to?

De Houwer suggests that the six properties we describe may reduce to a single feature, relations. Merely encoding relational contents is largely format-neutral – just as a smoke detector can encode propositions about the presence of smoke, an unstructured symbol like a lantern could, if embedded in the right sort of system, encode a proposition with relational content like <The British are approaching the shore of Massachusetts>. Of course we don't think this is what De Houwer has in mind by "relational content." De Houwer's pathbreaking work provides some of the strongest reasons for detecting LoT representations in implicit attitudes and "associative" learning (Mitchell, De Houwer, & Lovibond, 2009), and it's this sort of *explicit* representation of relations that implicates LoT structure. But we suspect that spelling out the notion of *explicit* relational content (beyond the mere representation of a relation) will require appealing to independently specified LoT properties – for example, predicate–argument structure (explicitly represented relations require multiple argument places) and role-filler independence (representing the same relations across distinct relata and vice versa). Thus we are skeptical that relational content is a more fundamental feature of LoT representations than those we specify or can play a role in grounding these other LoT properties that seem to us more fundamental. The hypothesis that relational contents *qua* multiplace predicates might constitute an important developmental and/or evolutionary advance in LoTs is an intriguing one, however, and we are grateful to De Houwer for raising the issue here.

Some commentators worried about the inference from LoT-like models to LoT-like structures in the mind. Griffiths, Kumar, & McCoy object that we "cross levels of analysis," and take LoT models at the Marrian computational level to support LoT structures at the Marrian algorithmic level (see also Roskies & Allen). They point out that deep neural networks (DNNs) can capture the inductive biases of Bayesian models without LoT structures. We agree that the successes of Bayesian models do not entail that the underlying mental representations share formal properties with the models. However, we draw no such inference in the target article. We do discuss computational models, but we do this (i) to point out that some computational models exploit LoT programs, and this undermines claims that the rise of DNN computational models has not made symbolic LoT approaches obsolete, and (ii) to note that the evidence that reaction time and error rates in encoding and searching for geometrical shapes tracks minimal description length in the probabilistic language-of-thought (PLoT), suggesting that the underlying algorithm implements this *specific* formal property (*viz.*, description length; Sablé-Meyer, Ellis, Tenenbaum, & Dehaene, 2021a). We don't take the modeling evidence to be decisive – instead, we use hundreds of experimental results to draw explanatory inferences about algorithmic-level representational structure. Therefore, although we grant the general point about the looseness between model and reality, we believe the target article takes pains to get at the underlying mental structures themselves.

McGrath, Russin, Pavlick, & Feiman (McGrath et al.) raise an important concern about the relationship between our six LoT properties and the LoT format itself. They are correct that we were unclear in our target article about whether these core properties are criterial or diagnostic of a deeper single cause. Our unclarity on this issue was not accidental – we are indeed unsure about the right answer. We have wondered whether LoT properties cluster together merely because that is what it is to

be a language (in a general, cognitively relevant sense). In other moods, we have been drawn to views where these properties cluster because they all subserve efficiency in domain-general reasoning, and perhaps are even necessary for domain-general computational systems²; we have also, in dark moods, been drawn to the idea that recursion is the true core of LoTH. However, in the end we just don't know what we think. Our working hypothesis has been that these properties are diagnostic rather than standard and there is a deeper reason why they cluster across systems and species.

We borrow the notion of homeostatic property clusters from philosophy-of-science, where it has been argued that instances of genuine natural kinds often³ share a common *mechanism* that explains the clustering of properties (Boyd, 1999; Craver, 2009). Perhaps, then, there is a mechanism underlying LoT phenomena (e.g., the still basically unknown formal or "syntactic" properties of LoTs), and these underlying mechanisms explain why the properties mentioned in our target article clump together. Nonetheless, facts like this are what you end up with at the end of inquiry, not the beginning. Without deciding between these views at the outset, we think the correct methodology is to search for a deeper fact that yokes these properties together. This is a big tent project and one that needs theorists from across the cognitive science spectrum. If it turns out that some of the properties we characterize are features of underlying cognitive mechanisms (e.g., discrete constituents) and others are emergent phenomena that these mechanisms produce (e.g., inferential promiscuity) then the target article will have missed out on key metaphysical facts about LoTs and the phenomena they generate. Some such possibility seems extremely likely to us. If our target article has outlined a strictly false but empirically useful characterization of an important kind of cognitive mechanism (a common way that science stumbles forward; Colaço, 2022), we would still consider that a success.

We agree with McGrath et al. that this is one of the most pressing questions at the core of the new iteration of LoTH that we offer. However, as they point out, the fate of LoTH as a viable hypothesis does not hang on the answer. It could be that the standard approach is best, and nonclassical architectures like DNNs could implement LoTs without underlying computational mechanisms that look especially symbolic. It could also be true that domain-general computation requires the cluster of LoT properties, such that any process that wanted to reach true formal computational power would have to end up with these properties one way or the other. If, on the contrary, there is some deeper mechanistic fact that causes LoT properties to cluster, a DNN could potentially instantiate them while lacking the underlying computational mechanism and thus be LoT-like, but not an instance of the same natural kind. Whether it is most fruitful to interpret LoTH as characterizing the underlying mechanism or the cluster of properties that can be produced by very distinct mechanisms is an open empirical question, and not one we need to answer in advance of using LoTH as a guiding hypothesis in cognitive science.

Chalmers's response marshals a similar distinction: Although we have argued for LoT representations, he argues, we are non-committal about the possibility of subsymbolic computation. On our version of LoTH, it is possible that "computational primitives (units) are not representational primitives."

Whether or not there is subsymbolic computation in biological cognition, there is also computation over LoT symbols. This

would be true even if, as he claims, “the quasi-symbolic operations of composition, decomposition, and quasi-logical inference may be available, but they are a tiny subset of the operations one can perform on the relevant distributed representations.” One reason why this is true is simply because compositional operations are *ipso facto* computations. At least four of the six features we describe concern the way LoT representations combine in thought to yield new representations. This is a form of computation. In this sense evidence for these features just is evidence for LoT-based computation, whether or not such computations can be implemented at the subsymbolic level, and whether or not subsymbolic computation also implements non-LoT operations.

Furthermore, a good deal of the evidence we cite concerns not merely compositional LoT sentences, but the use of those LoT sentences in cognition in real time. For example, the evidence surveyed in section 6 of the target article suggests that the logical form of LoT sentences is computed over automatically in system-1 reasoning; the evidence in section 4 of the target article suggests that the predicate–argument structure of LoT sentences is computed over when tracking objects; the evidence in section 5 of the target article suggests that physical reasoning computes over abstract content represented symbolically in LoT sentences and even logical representation of disjunction. All this evidence suggests that LoT sentences are not simply represented in the mind, but rather are used in computational processes that unfold across time. That is, *pace* **Aronowitz**, our evidence does not solely concern LoT representations “in stasis,” but rather illustrates that LoT representations are held in memory stores like VWM and figure in dynamical cognitive processes. And as **Antony** points out, LoTs are perfectly suited to make sense of reasoning at the personal level, bridging folk conceptions of mentality with a scientific one. We thank these commentators for their commentaries which allow us to foreground questions about the computational aspects of thought, which we agree is one of the most pressing issues in cognitive science.

It should not be surprising that the evidence for LoT structure tends to be evidence for LoT-sensitive computations – what good is an LoT if you can’t use it while thinking?

R5. Is LoTH generative?

R5.1. Vacuity

Several commentators have objected that our view is too slippery or vacuous to make for a good model of cognition (e.g., **Pereplyotchik**). **Attah & Machery**, for example, object that our pluralism about format undermines our defense of LoTH: Other formats could be doing the work that we attribute to LoTs, in particular in cases where we find evidence for only part of the property cluster. Logical space is full of possible representational formats, some of which are LoTs, others are not, and some are borderline and hard to categorize one way or the other. Some subset of these possible formats is instantiated in the minds of living creatures. Whether other, non-LoT representational formats explain the evidence we leverage from across different branches of cognitive science is an interesting empirical question, but one that could only be answered with careful attention to specific data. In particular, one would need to specify in detail the relevant features of the alternative formats for each case – formats that lack many of the six features we describe – and then provide evidence that they are doing the explanatory work, and not an

LoT. For now, we have made our proposal for how to explain this large amount of data, and we invite other researchers to show how and why they think LoTs do not offer the best explanation for specific cases.

Roskies & Allen raise a related objection: That our six properties are liberal to the point of vacuity – an “interpretative dance after the theory is already on offer” – because they allow for Treisman feature maps to be LoTs. We think feature maps are a prime example of a format that is not an LoT (see sect. 2 of the target article). But another reply to the criticism that our view is vacuous is simply to point to the robust empirical research program currently underway as detailed by the commentators who are carrying it out, to which we return now.

R5.2. Extensions

Jerry Fodor took his 1975 book to be merely collecting and codifying platitudes he saw in the research of cognitive scientists around him. And although we have distanced ourselves from certain features of his view, we very much share the idea that there has been something in the air already that we are simply tuning in to.

Yet as many of the commentaries have demonstrated, gathering broad commonalities may be helpful to cognitive scientists. One way they can be helpful is if the framework we sketch for identifying LoTs is extendable to cognitive systems where it has not yet been applied. For example, **Mahr & Schacter** fascinatingly use the features we describe to argue that episodic memory and imagination display LoT features, whereas **Cheng** explores the possibility of an LoT for touch.

Another way the framework can be helpful is through the creation or refinement of research programs. **Kibbe’s** commentary, for example, highlights the way our characterization of LoTH can be amenable to development research, despite the common assumption that they can’t, by helping developmentalists build testable hypotheses. In the same vein, **Demetriou** explores the possibility of a specific “Developmental LoT” and offers a hypothesis about how system-specific LoTs might develop over time. **Grüning** details methodological principles for studying LoT in social cognition in naturalistic settings. **Planer** offers an alluring, promising strategy for future work on the evolutionary origins of various LoTs using the sender–receiver framework.⁴ In vision, **Hafri, Green, & Firestone (Hafri et al.)** build on their seminal work on compositionality in vision, laying out a research program on a “‘psychophysics’ of compositional processes.” And **Westfall** looks at ways that LoT representations are at the front lines of artificial models of vision, complementing the cases we make for visual LoT and against DCNNs as models of biological vision. All of these views represent exciting areas for research we had not considered. We are deeply impressed by these commentators’ ingenuity.

Other areas for development include the theoretical foundations of LoTH, and we think two especially productive examples of this are the commentaries by **De Houwer** and **Antony**. De Houwer offers an alternative picture of the fundamental structures at play in LoTs (relations), whereas Antony complements our abductive case for LoT with an appeal to the explanatory need for LoTs to account for individual psychology.

These commentaries embody exactly what we had hoped would come of our target article: Clearly defined LoT-based research programs across the cognitive sciences, each developing in their own directions with proprietary debates and in some

cases, experimental details. We are extremely grateful for their commentators' contributions and excited to see how these programs develop in the coming years.

R6. Conclusion

A potted narrative of the history of cognitive science tells us that behaviorism died because Chomsky's review of *Verbal Behavior* killed it. But Skinner kept on doing what he was doing long afterward. We think it wasn't Chomsky's review, nor Festinger's work on cognitive dissonance (even though Festinger & Carlsmith, 1959, derive the exact opposite prediction from reinforcement theory), Milgram's on obedience, Miller's and Sperling's on memory, or any other specific findings/arguments. Arguments against views aren't generally why theoretical approaches fade away; it's usually that the theories decay when they no longer generate interesting questions. For example, Miller's work on memory caught the eye of a young woman who was wondering whether to work in biology, having moved on from anthropology. It wasn't the inconsistent and implausible features of behaviorism, but the fact that Miller's framework allowed her to ask specific, engaging, tractable questions about memory that drew her to study the mind. In this way, a generational talent turned to cognitivism, and because of that we get fast mapping, the theory-theory of concepts, bootstrapping approaches to concept acquisition, and the single most important force in the advance of developmental psychology (Carey, 2022). The trajectory of Carey is, we submit, not much different than that of the typical researcher – people are drawn to interesting questions that allow for some measure of progress, and shy away from recalcitrant theories that continue to spin their wheels.

A central role for theorists in cognitive science is to look at broad swaths of seemingly unrelated evidence and see if there is a thread tying those disparate research areas together. The target article attempted to do so by investigating areas where LoT would seem least likely, and offered six characteristics for an LoT that seemed to be mostly satisfied in all of these areas. A bigger picture emerged, on which the mind isn't an unstructured soup; rather it traffics in a certain format of thought and computation allowing for a common amodal code to subserve rational thought in areas that seemed less complex. We see the LoT offered here as an advance on the original theory, illustrating how our theories often under-intellectualize everyday cognition: Behind even the most seemingly reflexive, low-level areas of the mind lies a powerful, mechanistically rational computational engine.

Some theories are provocative, some productive. It is the rare theory that is both. What we aimed to do is show that LoTH is not only not dead, in fact it's currently one of the most fruitful theoretical frameworks in cognitive science. Nothing illustrates this point more than the inspiring commentaries we received – De Houwer, Demetriou, Dupre, Hafri et al., Kibbe, Mahr & Schacter, Planer, Wellwood & Hunter, and Westfall all show innovative new avenues to further the LoTH. There is no better evidence that LoT is the best game in town than looking at the incredible work that is being done in its name. Even people who disagree with us, such as Carey, Hochmann, McGrath et al., and Xu, do so in a way that forwards the empirical usefulness of the framework. We are grateful for their insights too. We may not be right in every detail; more importantly, by providing detailed theorizing we allow both our proponents and opponents places to do better research, and further insights into the working of the mind.

Notes

1. Nor do we agree that propositional contents should be thought of as worlds where these recognition procedures are successful. If the thought ALICE BEATS BART AT TUG-OF-WAR has a propositional content, it should turn out to be true in any world where Alice beats Bart at tug-of-war, even if it happens to be very foggy. What matters is that there was a tug-of-war game, Alice won, and Bart lost – because the recognition procedures can fail in all sorts of ways (Bart looks weird that day; it's dark out; etc.) while the proposition remains true, the two must be sharply distinguished. We relegated this point to an endnote because we're not sure if Oved et al. actually meant to suggest that “the proposition picks out the set of possible worlds where all those recognitions would happen.”
2. Thanks to Nick Shea for suggesting this possibility.
3. We note that Boyd allows for some properties in the distinctive cluster to “favor the presence of the others” (1999, p. 143) without explanation via underlying mechanism. McGrath et al.'s complaint that some of our properties favor the presence of others (e.g., predicate-argument structure and discrete constituents) may therefore fail to undermine the claim that the cluster constitutes a natural kind.
4. In the vein of efficient symbolic coding, as Planer mentions Gallistel's work, we should also mention a stunning paper by Akhlaghpour (2022) demonstrating the potential of RNA to function as the neurobiological basis of such coding.

References

- Akhlaghpour, H. (2022). An RNA-based theory of natural universal computation. *Journal of Theoretical Biology*, 537, 110984.
- Alderete, S., & Xu, F. (2023). Three-year-old children's reasoning about possibilities. *Cognition*, 237, 105472. <https://doi.org/10.1016/j.cognition.2023.105472>
- Block, N. (2023). *The border between seeing and thinking*. Oxford University Press.
- Boyd, R. (1999). Homeostasis, species, and higher taxa. In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 141–185). MIT Press.
- Camp, E. (2018). Why maps are not propositional. In A. Grzankowski & M. Montague (Eds.), *Non-propositional intentionality* (pp. 19–45). Oxford University Press.
- Carey, S. (2011). Précis of the origin of concepts. *Behavioral and Brain Sciences*, 34(3), 113–124.
- Carey, S. E. (2022). Becoming a cognitive scientist. *Annual Review of Developmental Psychology*, 4, 1–19.
- Carston, R. (2012). Word meaning and concept expressed. *The Linguistic Review*, 29(4), 607–623.
- Cavanagh, P. (2021). The language of vision. *Perception*, 50(3), 195–215.
- Cesana-Arlotti, N., Kovács, A. M., & Téglás, E. (2020). Infants recruit logic to learn about the social world. *Nature Communications*, 11(5999).
- Cesana-Arlotti, N., Martín, A., Téglás, A., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science (New York, N.Y.)*, 359, 1263–1266.
- Clarke, S. (2022). Mapping the visual icon. *The Philosophical Quarterly*, 72(3), 552–577.
- Colaço, D. (2022). What counts as a memory? Definitions, hypotheses, and “kinding in progress”. *Philosophy of Science*, 89(1), 89–106.
- Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, 22(5), 575–594.
- Engelmann, J., Völter, C. J., O'Madagain, C., Proft, M., Haun, D. B., Rakoczy, H., & Herrmann, E. (2021). Chimpanzees consider alternative possibilities. *Current Biology*, 31, R1–R3.
- Feiman, R., Mody, S., & Carey, S. (2022). The development of reasoning by exclusion in infancy. *Cognitive Psychology*, 135, 101473. <https://doi.org/10.1016/j.cogpsych.2022.101473>
- Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, 58(2), 203.
- Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford University Press.
- Fodor, J. A., & Lepore, E. (1998). The emptiness of the lexicon: Reflections on James Pustejovsky's *The generative lexicon*. *Linguistic Inquiry*, 29(2), 269–288.
- Green, E. J. (2023). The perception-cognition border: Architecture or format? In B. P. McLaughlin & J. Cohen (Eds.), *Contemporary debates in philosophy of mind* (pp. 469–493). Blackwell.
- Haladjian, H., & Pylyshyn, Z. W. (2008). Object-specific preview benefit enhanced during explicit multiple object tracking. *Journal of Vision*, 8(6), 497.
- Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, 24(2), 175–219.

- Leahy, B., Huemer, M., Steele, M., Alderete, S., & Carey, S. (2022). Minimal representations of possibility at age 3. *Proceedings of the National Academy of Sciences of the United States of America*, 119(52), e2207499119. <https://doi.org/10.1073/pnas.2207499119>
- Lin, Y., Stavans, M., & Baillargeon, R. (2022). Infants' physical reasoning and the cognitive architecture that supports it. In O. Houdé & G. Borst (Eds.), *Cambridge handbook of cognitive development* (pp. 168–194). Cambridge University Press.
- Lin, Y., Li, J., Gertner, Y., Ng, W., Fisher, C. L., & Baillargeon, R. (2021). How do the object-file and physical-reasoning systems interact? Evidence from priming effects with object arrays or novel labels. *Cognitive Psychology*, 125, 101368. doi:10.1016/j.cogpsych.2020.101368
- Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657), 279–281.
- Mandelbaum, E. (2014). Thinking is believing. *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, 57(1), 55–96.
- Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*, 50(3), 629–658.
- Mandelbaum, E. (2018). Seeing and conceptualizing: Modularity and the shallow contents of perception. *Philosophy and Phenomenological Research*, 97(2), 267–283.
- Mandelbaum, E. (2019). Troubles with Bayesianism: An introduction to the psychological immune system. *Mind and Language*, 34(2), 141–157.
- Mandelbaum, E. (2020). Assimilation and control: Belief at the lowest levels. *Philosophical Studies*, 177, 441–447.
- Mandelbaum, E., Dunham, Y., Feiman, R., Firestone, C., Green, E. J., Harris, D., ... Quilty-Dunn, J. (2022). Problems and mysteries of the many languages of thought. *Cognitive Science*, 46(12), e13225.
- Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, 32(2), 183–198.
- Murez, M., & Smortchkova, J. (2014). Singular thought: Object-files, person-files, and the sortal PERSON. *Topics in Cognitive Science*, 6(4), 632–646. <https://doi.org/10.1111/tops.12110>
- Murphy, E. (2021). Linguistic representation and processing of copredication (Doctoral dissertation), University College London, London.
- Mylopoulos, M., Pacherie, E., & Shepherd, J. (MS). The format of motoric representations.
- Odic, D., Roth, O., & Flombaum, J. I. (2012). The relationship between apparent motion and object files. *Visual Cognition*, 20(9), 1052–1081.
- Ortega-Andrés, M., & Vicente, A. (2019). Polysemy and co-predication. *Glossa: A Journal of General Linguistics*, 4(1), 1–23.
- Pepperberg, I. M., Gray, S. L., Cornero, F. M., Mody, S., & Carey, S. (2019). Logical reasoning by a grey parrot (*Psittacus erithacus*)? A case study of the disjunctive syllogism. *Behaviour*, 156, 409–445.
- Pietroski, P. M. (2018). *Conjoining meanings: Semantics without truth values*. Oxford University Press.
- Pomiechowska, B., & Gliga, T. (2021). Nonverbal category knowledge limits the amount of information encoded in object representations: EEG evidence from 12-month-old infants. *Royal Society Open Science*, 8(200782), 1–17.
- Porot, N. J. (2019). *Some non-human languages of thought* (Doctoral dissertation). CUNY Graduate Center.
- Porot, N., & Mandelbaum, E. (2020). The science of belief: A progress report. *WIREs Cognitive Science*, 12(2), e1539. <https://doi.org/10.1002/wcs.1539>
- Porot, N., & Mandelbaum, E. (2022). The science of belief: A progress report. In J. Musolino, J. Sommer, & P. Hemmer (Eds.), *The cognitive science of belief: A multi-disciplinary approach* (pp. 55–91). Cambridge University Press. doi:10.1017/9781009001021.005 (Reprinted with additions from *WIREs Cognitive Science*. 2020. <https://doi.org/10.1002/wcs.1539>)
- Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.
- Pylyshyn, Z. (2009). The empirical case for bare demonstratives in vision. In R. J. Stainton & C. Viger (Eds.), *Compositionality, context and semantic values: Essays in honour of Ernie Lepore* (pp. 254–274). Springer.
- Quilty-Dunn, J. (2020c). Perceptual pluralism. *Noûs*, 54(4), 807–838.
- Quilty-Dunn, J. (2021). Polysemy and thought: Toward a generative theory of concepts. *Mind & Language*, 36, 158–185.
- Quilty-Dunn, J., & Green, E. J. (2023). Perceptual attribution and perceptual reference. *Philosophy and Phenomenological Research*, 106(2), 273–298.
- Quilty-Dunn, J., & Mandelbaum, E. (2018). Against dispositionalism: Belief in cognitive science. *Philosophical Studies*, 175, 2353–2372.
- Quilty-Dunn, J. (forthcoming). Sensory binding without sensory individuals. In A. Mroczko-Wasowicz & R. Grush (Eds.), *Sensory individuals, properties, & perceptual objects: Unimodal and multimodal perspectives*. Oxford University Press.
- Sablé-Meyer, M., Ellis, K., Tenenbaum, J., & Dehaene, S. (2021a). A language of thought for the mental representation of geometric shapes. *PsyArXiv*. doi:10.31234/osf.io/28mg4
- Shepherd, J. (2019). Skilled action and the double life of intention. *Philosophy and Phenomenological Research*, 98(2), 286–305.
- Srinivasan, M., & Rabagliati, H. (2015). How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua. International Review of General Linguistics. Revue Internationale de Linguistique Generale*, 157, 124–152.
- Stavans, M., & Baillargeon, R. (2018). Four-month-old infants individuate and track simple tools following functional demonstrations. *Developmental Science*, 21, e12500.
- Stavans, M., Lin, Y., Wu, D., & Baillargeon, R. (2019). Catastrophic individuation failures in infancy: A new model and predictions. *Psychological Review*, 126(2), 196–225.
- Taylor, J., & Xu, Y. (2021). Joint representation of color and form in convolutional neural networks: A stimulus-rich network perspective. *PLoS ONE*, 16(6), e0253442.
- Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14(1), 107–141.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136.
- Vicente, A. (2018). Polysemy and word meaning: An account of lexical meaning for different kinds of content words. *Philosophical Studies*, 175(4), 947–968.
- Westfall, M. (forthcoming). Perceiving agency. *Mind & Language*.
- Wolfe, J. M. (2021). Guided Search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, 28(4), 1060–1092.
- Xu, F. (2019). Toward a rational constructivist theory of cognitive development. *Psychological Review*, 126(6), 841–864.
- Xu, F., & Carey, S. (1996). Infants' metaphysics: The case of numerical identity. *Cognitive Psychology*, 30(2), 111–153.