The Language of Thought Hypothesis as a Working Hypothesis in Cognitive Science[1]

Jake Quilty-Dunn, Washington University in St. Louis, USA, quiltydunn@gmail.com, sites.google.com/site/jakequiltydunn/

Nicolas Porot, Africa Institute for Research in Economics and Social Sciences, Mohammed VI Polytechnic University, Morocco, nicolasporot@gmail.com, nicolasporot.com

Eric Mandelbaum, The Graduate Center & Baruch College, CUNY, USA, eric.mandelbaum@gmail.com, ericmandelbaum.com

Abstract:
The target article attempted to draw connections between broad swaths of evidence by noticing a common thread: abstract, symbolic, compositional codes, i.e., 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 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.

1. LoT and Development

The topic of development arose frequently, raised by **Canudas-Grabolosa, Martín-Salguero, & Bonatti; 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

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

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). Though 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 Piagettian 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 non-monotonic 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 pre-linguistically, 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. While 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 three years old sequentially simulate possibilities in 3- and 4-cup tasks (Leahy et al. 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 et al. 2018), as can the holding of multiple alternatives in mind at once on a task similar to the 2-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 2-cup task failures for 14- and 15-month olds (Feiman et al. 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 3-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? 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 due to 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: e.g., 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 & 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 & 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 et al. 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. Since 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. Since 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 which 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 a LoT (Quilty-Dunn & Mandelbaum 2018, Porot & Mandelbaum 2020/2022). 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 & Baillargeon (2018), Stavans et al. (2019), and Lin et al. (2021) results. Children's failures to use basic-level categories for object individuation before 12 months might be due to 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 forthcoming). 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 & 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, 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 & Carey (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.

2. 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, Lupyan**, **Xu**), some worried that the argument overextends (**Cheng**, **Roskies & Allen**, **Attah & Machery**), and some saw opportunity for testable claims about development (**Carey**, **Kibbe**, **Xu**), which we discuss in Section 1. 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 due to 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 et al. 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 non-discrete 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, 116) and Xu (2019) have conceived of the idea enough to argue explicitly against it, and Pylyshyn (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, since 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 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 Section 5 below) 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 due to 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; i.e., 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 visual working memory. 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 a 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.

3. 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 "crystallization" 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 non-human LoTs might differ syntactically from any known natural or formal languages. Similarly, perhaps some syntactic features of natural language differ from those of pre-linguistic 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 Section 1 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, since 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 re-emerged 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 a LoT.

We agree with **Oved, Krishnaswamy. Pustejovsky, & Hartshorne** 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—e.g., how LoTs interact with images, analog magnitudes, motor intentions (Shepherd 2017; Mylopoulos et al. MS), 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. While 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.[2] 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 anti-pragmatist view we are drawn to.

Oved et al. also raise the problem of polysemy, i.e., 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,

[2] 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—since 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 a footnote 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".

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 co-predication ("Lunch was delicious and informative") (Murphy 2021). Unlike homonymy, forms of so-called "regular" (i.e., systematic and rule-governed) polysemy—like, e.g., 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 a 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 non-atomic concepts, pieces of which can be deployed on different occasions (Ortega-Andres & 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 paper 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 non-human 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. Since Kaufmann &Newen did not discuss these paradigms or domains, we are left wondering which aspects of our application of LoTH to non-human 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 non-human animals.

Antony points to applications of LoT to person-level phenomena involving beliefs and other propositional attitudes. We wholeheartedly agree. While 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 anti-representationalism 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; Mandelbaum 2016; Quilty-Dunn & Mandelbaum 2018; Porot & Mandelbaum 2020). 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. 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.

4. 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 et al. 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—e.g., 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* multi-place 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 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 PLoT, suggesting that the underlying algorithm implements this *specific* formal property (viz., description length; Sablé-Meyer et al. 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, while 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 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 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 systems;[3] 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 criterial 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 often[4] 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 paper 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 criterial approach is best, and non-classical 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 another. If, on the other hand, there is some deeper mechanistic fact that causes LoT properties to cluster, a DNN could potentially instantiate them while lacking the underlying

[3] Thanks to Nick Shea for suggesting this possibility.

[4] We note that Boyd allows for some properties in the distinctive cluster to "favor the presence of the others" (1999, 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.

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' response marshals a similar distinction: While 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 suggests that the logical form of LoT sentences is computed over automatically in System 1 reasoning; the evidence in Section 4 suggests that the predicate-argument structure of LoT sentences is computed over when tracking objects; the evidence in Section 5 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 visual working memory 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 authors 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 a LoT if you can't use it while thinking?

5. Is LoTH Generative?

5a.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 and 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 a 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 a LoT (see Section 2). 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.

5b. 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 while 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 & Schachter** fascinatingly use the features we describe to argue that episodic memory and imagination display LoT features, while **Cheng** explores the possibility of a 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.[5] In vision, **Hafri, Green, and Firestone** build on their seminal

[5] 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

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 authors' 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), while 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 authors' contributions and excited to see how these programs develop in the coming years.

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 afterwards. We think it wasn't Chomsky's review, nor Festinger's work on cognitive dissonance (even though Festinger and 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 Susan 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 6 characteristics for an LoT which 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 trafficks 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 & Schachter, 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.**, & **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.

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.

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.

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.

Feiman, R., Mody, S., & Carey, S. (2023). 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: 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 Brian P. McLaughlin & Jonathan Cohen (eds.), *Contemporary Debates in Philosophy of Mind* (Oxford: Blackwell), 469-493

Haladjian, H. & Pylyshyn, Z.W. (2008). Object-specific preview benefit enhanced during explicit multiple object tracking. *Journal of Vision*, 8(6), 497.

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*, 119(52), e2207499119. https://doi.org/10.1073/pnas.2207499119

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*, *57*(1), 55–96.

Mandelbaum, E. (2020). Assimilation and Control: Belief at the Lowest Levels. *Philosophical Studies*, *177*, 441-447.

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, UCL (University College 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).

Pietroski, P. M. (2018). *Conjoining meanings: Semantics without truth values*. Oxford University Press.

Porot, N., & Mandelbaum, E. (2020). The science of belief: A progress report. *WIREs Cognitive Science*. https://doi.org/10.1002/wcs.1539

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Quilty-Dunn, J. Sensory binding without sensory individuals. In Aleksandra Mroczko-Wasowicz & Rick Grush (eds.), *Sensory Individuals, Properties, & Perceptual Objects: Unimodal and Multimodal Perspectives* (Oxford: Oxford University Press).

Quilty-Dunn, J., & Mandelbaum, E. (2018). Against dispositionalism: Belief in cognitive science. *Philosophical Studies*, *175*, 2353-2372.

Shepherd, J. (2019). Skilled Action and the Double Life of Intention 1. *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*, *157*, 124-152.

Taylor & Xu 2021 Joint representation of color and form in convolutional neural networks: A stimulus-rich network perspective

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

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive psychology*, *14*(1), 107-141.

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. (2022). 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.