Received: 12 March 2021 | Revised: 12 August 2021 | Accepted: 6 October 2021

DOI: 10.1111/mila.12400

SUBMITTED ARTICLE

WILEY

Public language, private language, and subsymbolic theories of mind

Gabe Dupre

School of Social, Political, and Global Studies, Keele University, Staffordshire, UK

Correspondence
Gabe Dupre, School of Social, Political, and Global Studies, Chancellor's Building, Keele University, Staffordshire ST5 5BG, UK.
Email: g.g.dupre@keele.ac.uk

Funding information
Funding for this research was provided by the Leverhulme Trust.

Language has long been a problem-case for subsymbolic theories of mind. The reason for this is obvious: Language seems essentially symbolic. However, recent work has developed a potential solution to this problem, arguing that linguistic symbols are public objects which augment a fundamentally subsymbolic mind, rather than components of cognitive symbol-processing. I shall argue that this strategy cannot work, on the grounds that human language acquisition consists in projecting linguistic structure onto environmental entities, rather than extracting this structure from them.

KEYWORDS
bootstrapping, connectionism, developmental linguistics, extended mind, generative linguistics, symbolic cognition

1 | INTRODUCTION

Contemporary theoretical psychology and philosophy of cognitive science offers a bewildering array of contrasting positions. Theories can be nativist or empiricist, classical or connectionist, extended or head-bound, and so on. If adopting a position requires settling all such questions, one may wonder: Who has the time? Fortunately, in practice, these debates are not independent. Answers to one tend to motivate decisions with respect to others. Sociologically, this has led to theorists clustering loosely into two broad "camps." On one side, there are "traditionalists," exemplified by Jerry Fodor and Noam Chomsky, who adopt classical, symbolic-systems approaches to cognition, with significant innate structure, and tend to abstract away from the mind's bodily and environmental context. On the other, there are

"radicals," such as Andy Clark, who propose subsymbolic systems, with minimal substantial innate structure, which are augmented, scaffolded, and even composed, by extracranial relations.[1]

This clustering has several consequences for theoretical investigation of the mind. On the one hand, it may make such investigation more efficient, in that finding compelling reasons to adopt a stance with respect to one of these high-level debates can push one toward others. On the other, the answers within the cluster can be mutually reinforcing. In some cases, what looks like a problem for a position in one of these debates can be resolved by the other views in that cluster. I shall describe a case study of this sort of reasoning.

Language provides a paradigmatic case for the traditionalists. That humanity's most distinctive psychological capacity is a capacity to use a symbolic system has seemed to provide a strong case that, in at least some core respects, the human mind is a symbolic system. However, a wide range of theorists, from both psychology and philosophy, have argued that this suggestion is misleading. Classical theorists, it is argued, have abstracted away from the ways in which the use and acquisition of language is essentially a matter of engaging with *public symbols*. When the publicity of language is ignored, it may seem that linguistic competence must consist in internal, psychological, symbol-manipulation. But once the externality of language is taken seriously, the symbolic nature of language can be kept outside the mind, and the subsymbolic approach to psychology can be retained.

Clark (1998b) presents perhaps the clearest statement of this proposal: "The coding system of public language is thus especially apt to be co-opted for more private purposes" (p. 178). Clark develops this proposal extensively (see, e.g., Clark, 1996, 2006; Clark & Karmiloff-Smith, 1993) but similar views are expressed by Lupyan and Bergen (2016) and Lupyan (2012) in the claim that language "programs" the mind, by Heyes's (2018) claim that language is a "cognitive gadget," Rumelhart et al.'s (1986) description of language as "a kind of internalization of an external representational format" (p. 47), and Tomasello (2019), whose "Neo-Vygotskian" theory of development echoes the claim that language is "internalized" on the basis of public symbols.

The core idea here is that a subsymbolic (e.g., connectionist) mind genuinely cannot reproduce the behavior of a symbol-manipulating system. However, this only shows that *our grasp* of language is subsymbolic, it does not show that language itself is. By locating language *mind-externally*, as a feature of our constructed environment, we can reconceptualize the cognitive process of language use and acquisition. Linguistic capacities are capacities to interact with a genuinely symbolic, but public, system, not to internalize such a system. Language is thus better conceived as a cognition-enhancing artifact, rather than an intrinsic component of the mind itself. If this is correct, the failure of subsymbolic systems to account for our linguistic capacities indicates not a limitation of these approaches, but instead a misunderstanding of in what these capacities consist. Theories of the mind should not be burdened with explaining how the symbolic structure of a language can be acquired and utilized any more than anatomical theories should be expected to explain airplane-assisted flight.

I begin this article by spelling out be the core difference between symbolic and subsymbolic theories of mind, namely the presence of repeatable symbols. I then show why linguistic competence presents a deep worry for subsymbolic theories. In Section 4, I will present the "off-loading" proposal just discussed, according to which language should be viewed not as a

[1]Intermediate positions are possible, indeed plausible. But these two approaches define poles between which we can locate proposals for theories of the structure of the mind.

symbolic system of mind, but instead as a symbolic public system with which the subsymbolic mind can interact. In Sections 5 and 6 I present my case that this maneuver cannot achieve what its proponents want it to, on the grounds that language acquisition is a matter of *imposing* linguistic structure on public stimuli, not a matter of "internalizing" an antecedently available system, first presenting the case generally, then specifically discussing "bootstrapping" models of acquisition suggested in the literature.

2 | MENTAL SYMBOLS

"Classical" approaches to cognition explain cognitive capacities with reference to the rule-governed manipulation of discrete representational symbols. Paradigmatic mental processes involve the generation of such symbols on the basis of interaction with the environment (sensation), interaction with the environment on the basis of such symbols (action), and the generation of symbols on the basis of other symbols (reasoning). The visual system, for example, can be understood, according to Marr (1982), as functioning to generate a series of representations of the local environment, in response to sensory stimulation. On this classical approach, the job of cognitive science is to identify the stock of such symbols available to the system and the rules governing the production and manipulation of such symbols.

A rival approach, connectionism, is widely favored in many branches of cognitive theorizing. In opposition to the abstract, flow-chartesque systems of the classical approach, connectionist models are rooted in dynamic models of neural activation patterns. Connectionist models consist of (usually large) collections of interconnected nodes. Input nodes transduce environmental stimuli, output nodes exhibit the results of the processing of the system, and hidden nodes perform the processing connecting the former to the latter. Each connection is associated with a weight, which determines how the firing of one node influences that of all the other nodes immediately downstream of it. The firing of the output nodes is thus determined by that of the hidden nodes, in accordance with the latter's weights, and that of the hidden nodes is likewise determined by the firings, and weights, of the input nodes.

The crucial feature of these networks is their ability to evolve over time, so as to improve at whatever task they are assigned. In a simple example a network tasked with classifying images could be constructed with each input node assigned to some region of the input images, and the output nodes assigned to several outputs each corresponding to a classification. Improving at this task would depend on determining which features of the input provide the best evidence for one output or another. If the system begins with the weights between each node generated randomly, it would perform at no better than chance. However, if early performance is monitored and compared with the correct classification, it can learn from its mistakes, changing these weights so as to make it more likely to get things right. Complex connectionist systems, trained on large datasets, have demonstrated cognitive abilities such as visual item recognition, sentence parsing, and more. These successes have encouraged those who think connectionist networks may provide the key to understanding the mind generally.

For our purposes, the crucial difference between these two models of mental processing is that in classical models, the entities over which such processes are defined, mental symbols, are *repeatable*. That is, tokens of the same mental type can serve as constituents of distinct mental states or processes. The paradigmatic example comes from propositional thought, in the claim that there is some psychological token which is *shared* by the beliefs that LeBron James is tall and that LeBron James is fast. In order to form such thoughts, a classical approach to mind

insists that one produces two tokens of the same mental symbol representing LeBron James. The semantic similarity of these two thoughts (i.e., that both predicate something of the same person) is explained by the (qualitative) *identity* of their syntactic constituents.

Connectionist systems, on the other hand, do not require that there be anything corresponding to such repeatable symbols. Mental processes, on such an account, are patterns of activation of nodes within a neural network. Semantically similar states may exhibit *similar* patterns of activation, but there will rarely be any way of *identifying* subcomponents of these states, in the way we could in classical systems.

Smolensky (1991) provides a standard example, in discussing the ways a simple connectionist system might represent a cup of coffee. The classical paradigm has it that such a representation would feature symbols for CUP and COFFEE, and that these would be the same symbols used in representing a cup of water or a pot of coffee. Indeed, when given the representation of a cup of coffee, we could identify the repeatable representation COFFEE by subtracting the structure for A CUP OF ——. Smolensky claims that connectionist networks need not have this property. When connectionist representations are combined, there may be interaction effects which preclude such constituent identification. In our CUP OF COFFEE network, subtracting those features characteristic of CUP representations does not lead to a context-invariant COFFEE symbol. Rather, there are traces of its combination. For example, the remnant network will have something in common with the network's representations of other liquids of similar shapes. Smolensky describes this as an overlapping set of nodes, each corresponding to a semantic feature. For example, COFFEE, in the context of CUP OF COFFEE will share the node BROWN LIQUID WITH CURVED SIDES AND BOTTOM which would likewise be activated as part of the network's system of representing CUP OF BOURBON and even MUDDY PUDDLE.[2]

Connectionist networks, then, do not have constituent structure in the way classical networks do, as the components of complex networks are not repeatable. Complex representations do not, really, have simpler ones as constituents. Instead, they are products of the interactions of these simpler representations. In contrast to the building-block-like structures of classical systems, in which simple components of complex structures retain their properties and independence, connectionist processing is more like paint mixing: There will typically be similarities between the inputs and the products, but the former cannot be found as parts of the latter.

As in any interesting high-level scientific dispute, each proposal has domains in which it applies more naturally. Connectionist models are apropos for cognitive tasks requiring graded classification, especially when this depends on identifying a large number of cues which are individually of minor significance. The complex structure of these networks, and the ability to readjust connection weights (pseudo-)continuously, enables them to identify these subtle patterns robustly and reliably. While such capacities and properties are not precluded by a classical architecture, they do not emerge so naturally.

However, when a task involves genuinely categorical inference, that is, when a given input requires a specific output regardless of the surrounding context, classical systems are better equipped. Imagine constructing a system which, whenever COFFEE was tokened, responded by tokening TEA. In a classical system, this is, of course, trivial. Rules governing transitions from one representation to another are the very heart of such systems. However, it is not clear how

[2]In Smolensky's model system, we can identify repeatable components of these complex representations: the nodes. While CUP and COFFEE are not repeatable, the nodes composing them (BROWN, LIQUID, etc.) are. But this is not an essential feature of connectionist networks, as the nodes themselves may not have any context-invariant meaning. See Clark (1993, Chapter 2) for detailed discussion.

to even state such a rule in a connectionist system composed of distributed, nonrepeatable representations. In such systems, the "tokening of COFFEE" does not correspond to a distinctive set of events, but rather a cluster of patterns of activation which resemble one another (but also other, nontarget, states) in certain ways.

In this way, a lack of repeatable symbols means that connectionist systems can instantiate rules of the above sort only in an approximate and often accidental sense. These systems approximate these rules in the sense that describing the system's states as COFFEE or TEA representations does not classify the states of the system at the level of grain that the system itself does. The sensitivity of the systems is much more fine-grained than this. Thus, describing these systems in these terms is a coarse, but efficient, way of characterizing the patterns displayed by the system, rather than a way of referring to genuine, causally significant features of the system.

The realization of rules is accidental, in the sense that the nature of the system does not guarantee it will be instantiated, but may depend on the tokens that happen to be realized. While it may be that, for any given system, all tokenings of COFFEE are indeed followed by tokenings of TEA, this fact does not guarantee that this will hold for all possible tokenings. In novel contexts, COFFEE representations will have different features (e.g., in the context —— ON THE TREE, the representation of COFFEE may be more similar to BRAZIL NUT). While the system may be (arbitrarily) good at projecting to future contexts, without a repeatable symbol there is no guarantee that future tokenings will lead to the same responses.

This is not merely a conceptual possibility. Various subsymbolic models of language acquisition have been produced which initially seem to replicate the rule-like nature of linguistic competence, only to be shown to fail when applied to a broader dataset, thus confirming that the instantiation of rules was indeed accidental and approximate. Pinker and Prince (1988) present a famous case of this, responding to Rumelhart and McClelland's (1986) claim to have shown that a connectionist network was capable of learning English past tense, by applying this model to unusual verbs, for which these networks produce bizarre results ("smeej" becomes "leefloag," rather than "smeejed"). More recently, Kam et al. (2008) showed that Reali and Christiansen's (2005) claims to have shown that purely statistical analyses of linguistic input are sufficient to learn that question-formation in English is structure-dependent fail in analogous ways: When these models are applied to a wider dataset, the approximation of rules fails. Kam et al. showed that the statistical properties relied on in Reali and Christiansen's (admittedly very simple) model to differentiate grammatical ("Is the little boy who is crying hurt?") from ungrammatical ("Is the little boy who crying is hurt?") sentences depended on contingent features of the constructions in question and thus did not generalize in the way that genuine grammatical knowledge should. Specifically, the statistical model identified sentences of the former type as grammatical on account of containing the bigram "who/that is," which was found frequently in the corpus, while the latter was ungrammatical on account of containing "crying is," which was not attested. But the frequency of the bigram "who/that is" was a product entirely of irrelevant linguistic constructions like wh-questions ("who is that?") and demonstratives ("that is a horse"). Given that not all yes–no questions in English feature this bigram and that cross-linguistically the homophony between the relative particles "who/that" and question and demonstrative pronouns is not robust, we cannot view this model as genuinely learning how question-formation works, but simply as identifying a contingently useful cue for grammaticality. When this cue was eliminated, model performance dropped drastically.

These results exemplify exactly the accidental and approximate nature of such statistics-based systems. That is, the approximation of a rule or inference by a subsymbolic system holds

not in virtue of the categories with which we abstractly describe the rule, but instead by the fine-grained associations between all the various nodes relevant to the classification of the system as exemplifying this rule. An inference from COFFEE to BROWN will be realized on the basis not of some significant connection between COFFEE and BROWN, but in virtue of myriad associations between the finer-grained properties that serve to realize the system's approximations of COFFEE or BROWN on a given occasion. This difference is thus analogous to the difference between laws, which hold in virtue of connections between the properties themselves ("all noble gasses are inert") and accidental generalizations, which hold in virtue of properties that particular tokens of these properties happen to have ("all my siblings are vegetarian").[3]

These points are related to the fact that in classical systems, there is a clear distinction between rules and the symbols over which they are defined. However, in connectionist systems, no such boundary can be drawn. While theorists can describe connectionist systems as containing representations and instantiating rules, these are both more-or-less rough abstractions over the spreading patterns of activity in the network, with representations corresponding to momentary patterns of activation, and rules corresponding to causal-temporal relations between these. This reverses the mode of explanation from the classical approach. Classically, the representations and the rules governing them constituted the *explanans*, and patterns of behavior were explained with appeal to these. In connectionist systems, however, representations and rules are little more than summaries of the genuinely causal and explanatory connections and weights.

This section has hopefully shown why a connectionist system will have difficulty with processes which involve the application of rules. Namely, without an ability to repeat type-identical symbols, there can be no guarantee that triggers for a rule will always be recognized as such. Rule-governed processes may be approximated by such systems, in somewhat accidental fashion, by state transitions which respond to similar-enough antecedents, but if we have reason to view some aspect of the mind as rule-governed in a more robust sense this may be insufficient.

Proponents of connectionist systems (e.g., Christiansen & Chater, 1999; Clark, 1993; Smolensky, 1988) typically motivate them by stressing that they are not rigidly rule-governed in the way classical systems are. Many of the desirable properties of these systems stem from this fact: graceful degradation, application to fuzzy categories, continuity with lower-level cognition, plausible neural implementation, and so on. I will thus continue to take this lack of repeatable symbols, and thus lack of rules defined over such symbols, as criterial of connectionist systems. To the extent that rules and repeatable representations are found in such systems, I will take them to be implementations of classical systems, and thus outside the scope of my challenge.

3 | THE PROBLEM OF LANGUAGE

Fodor and Pylyshyn (1988) famously argued that rules and representations must feature in a cognitive architecture, for they are needed to explain the systematicity and productivity of thought and language. My argument agrees with this conclusion, but need not accept the reasoning to get there. That human language is systematic, and rule-governed, is simply an empirical result. The most compelling and best confirmed theories of human language treat it as a system of rules. This is paradigmatically true in syntax, but also phonology (Berent, 2013; Hale & Reiss, 2008), morphology (Prasada & Pinker, 1993), and more-or-less all formal

[3]See van Fraassen (1989, p. 27) and Dretske (1977).

approaches to semantics. Thus, whether Fodor and Pylyshyn's transcendental arguments are successful or not, I take our linguistic capacities to pose distinctive problems for any subsymbolic approach to mind.

Consider the way that functional categories such as negation, modality, tense, and aspect interact with verbal inflection and word order in English. Every English sentence must be marked for tense in its main clause, but this marking is not provided by standalone lexical items, but instead as inflection on a verb. In simple cases, this tense marking appears on the lexical verb (1). However, when the sentence contains an auxiliary verb as well, this, rather than the lexical verb, serves as the target of tense-marking (2). Marking the lexical verb as well results in ungrammaticality (3). We can introduce negation (4), which is always interposed between the auxiliary and the lexical verb.

1. I ran.
2. I could run.
3. *I could ran.
4. I could not run.
5. I did not run.
6. *I did not could run.

This paradigm suggests a general structure for English declaratives: Subject–tense–(negation)–verb. As the tense marking cannot stand on its own, when negation is absent, it is phonologically realized by affixation to the adjacent verb. But when negation is present, this creates a distance between the tense marker and the verb, blocking affixation. This leaves the tense particle stranded, without a suitable host. There are various options for salvaging this defective structure. One is to reshuffle the constituents, moving the verb over the negation, making it available for affixation. This option is taken in French ("*Je (ne) cours pas*") and Old English. Lexical verbs do not move in this way in English, however. But auxiliary verbs do, as seen in (4). In cases like (4), the auxiliary is there for independent reasons, but this need not be the case. To negate simple sentences like (1), without a semantically motivated auxiliary and without moving a lexical verb, we thus insert a "dummy" auxiliary (usually, "do") which can be moved to a position where it can host the tense marking of the sentence (5).

Explanations of this sort, paradigms of linguistic theory since Chomsky (1957/2020), rely essentially on the linguistic system's ability to classify linguistic expressions in repeatable, abstract ways. English speakers recognize that they must apply an additional rule ("'do"-insertion) when forming a negative sentence without a semantically motivated auxiliary verb, and that they must not apply this rule when there is such an auxiliary verb present (6). This categorization is abstract in that it is not reducible to observable features of the expressions in question. This is evidenced by the differential treatment given to homonymous expressions, such as "can" and "have." When these are lexical verbs, rather than auxiliaries, they are prohibited from movement for tense marking:

7. I can tuna (in a cannery).
8. *I can not tuna.
9. I do not can tuna.
10. I have three siblings.
11. *I have not three siblings.
12. I do not have three siblings.

Further, the distinction between auxiliary and lexical verbs is significant across arbitrarily different contexts. English speakers recognize that (13) and (14) are good while (15) and (16) are bad even though the crucial difference is deeply buried in grammatical structures they are unlikely to have had much experience with before.

13. I told my partner about the time my professor told me that I might not pass the exam.
14. I told my partner about the time my professor told me that I did not pass the exam.
15. *I told my partner about the time my professor told me that I passed not the exam.
16. *I told my partner about the time my professor told me that I not passed the exam.

These differences are easy enough to account for in a system with rules defined over repeatable symbols. The rule can simply refer to the abstract grammatical types *lexical verb* and *auxiliary verb*. These types form equivalence classes of lexical items, each member of which can then be treated in the same way by the rule. Once the system classifies a verb as lexical, it is prohibited from movement for tense-marking, no matter how superficially similar it is to auxiliary verbs (as in "can" and "have") or what context it is found in.

Connectionist systems, however, are (by definition) unable to appeal to this style of explanation. Even if we allow that such systems represent expressions *as* auxiliary or lexical verbs, they do so only in the sense that token representations of these sorts are more-or-less similar to one another. As similar patterns of activation cause similar downstream behavior, this means that such systems can imitate rule-based systems, within a particular range, allowing for a limited degree of projection to novel cases. But they cannot guarantee that some future token will be treated in the same way. The graded notion of similarity on which our attribution of representational categories to such systems relies ensures that significant enough variation along this axis will lead to a different style of response. On the one hand, if the auxiliary "have" starts looking too similar to the lexical verb, it may project sentences like (11). On the other, arbitrarily complex grammatical structures may undermine the contextual cues needed to classify appropriately, leading to failures such as (15) or (16).

This should not be controversial. There is no conceptual space between an inability to utilize repeatable, invariant symbols and the failure to guarantee that novel cases of a kind will be treated in the same way as previous instances. If representational kinds are mere abstractions over the real, causally significant, subsymbolic action, then being members of the same representational kind cannot ensure identical responses. But in the linguistic example just discussed, the capacity to generalize over arbitrary novel linguistic expressions is precisely what our best linguistic theories posit. So, there seems to be at least one psychological system, language, which connectionist systems cannot explain.

The point here is not that subsymbolic systems will necessarily make incorrect predictions about observable linguistic behavior. While it is true that all such systems have so far made such bad predictions, this is of course true for any rule-based system as well. The point is rather that, to use the distinction drawn by Bogen and Woodward (1988), such systems are not adequately describing the *phenomenon*. As the linguistic theories just discussed witness, viewing language as a system for the recursive application of rules to repeatable symbols makes for better explanations of linguistic observations. As has long been noted (see Chomsky, 1966), the ability of language users to project language in novel, but constrained, ways is a central explanandum for theoretical linguistics. And a rule-governed system seems to capture and explain this creativity better than a subsymbolic one. While we may recognize that we would not be able to parse a 1000-word long sentence, competent English speakers seem capable of

judging that such a sentence's grammaticality would depend on the same sorts of factors as those of smaller, interpretable sentences. Even if the main lexical verb of such a sentence is buried within nested clauses opaque to human perceptual systems, we can say confidently that it should not be raised above a negation for tense-marking, whereas an auxiliary verb may be. Such facts cannot easily be brought out in observational data, but it is the firm belief of most linguists, across all branches of theoretical linguistics, that they comprise the phenomenon to be explained nonetheless.

This point relies on the distinction between competence and performance. While performance, linguistic behavior including intuitions/judgments, comprises the observational data of linguistic theory, it is not the phenomenon to be explained. That is competence, the underlying capacity. As finite performance will of necessity deviate from competence, observations alone will not determine what our theories are supposed to explain. For that, a theoretically motivated judgment about what the significant phenomena are is needed. And I follow mainstream linguistics in judging that the underlying competence consists in a rule-governed computational system.

It is, at this point, open to defenders of broadly connectionist outlooks to reject such approaches to language. Christiansen and Chater (2016) ("C&C") do just this, relying on the fact that there is nothing in the observational data that forces one to accept that language is genuinely rule-governed in the way just described. As any sample of human linguistic behavior will necessarily be finite, it is always possible for a subsymbolic system to accurately approximate such behavior. C&C rightly recognize that the rejection of a symbolic approach precludes them from explaining purported linguistic facts such as that "Bulldogs bulldogs bulldogs fight fight fight" is a grammatical sentence of English (pp. 232–234). Given the rarity of a string of three instances of a single lexical item, it is highly improbable that any system which did not treat grammar in an abstract, rule-governed way would license a sentence like this. For this reason, C&C, rather than accept a gap here between performance and competence, simply accept that this sentence is *not* in fact a grammatical English sentence, just as "unaided syntactic intuitions" (p. 232) would have it. They argue instead that the form of reasoning that leads us to think that it is, despite these intuitions, is a "complex, meta-linguistic" "extension" of our linguistic competence.

The problem with this argument is that it misunderstands what such "meta-linguistic" reasoning shows. On C&C's telling, the fact that ordinary speakers are capable of such reasoning is the phenomenon to be explained. And they offer an extra-linguistic account of what such reasoning consists in. This is to treat the judgment that "bulldogs bulldogs bulldogs fight fight fight" is grammatical, made after one has worked through the metalinguistic argument to this effect, as a piece of performance data to be explained. But what is actually going on here is that the theorist is pointing to some uncontroversial facts about linguistic competence, and showing that they entail another *fact about competence*. Specifically, in English, it is uncontroversial that an object-relative clause can be formed by moving the object of a complete, transitive sentence to the beginning of the sentence. From this fact it simply follows, without stipulation to the contrary, that sentences like the bulldog sentence are part of English. This would be the case even if we *could not*, by meta-linguistic reasoning or otherwise, convince ordinary speakers of this fact. That is to say that this is a fact about our competence, however it plays out in performance.

Given the nature of subsymbolic systems, it will always be possible to arbitrarily closely approximate performance. But this is not, and should not be, the goal of linguistic theorizing. The goal is instead to describe the system that, in concert with a host of other capacities, makes

such performance possible. For this reason, I shall assume for the remainder of the article that mainstream linguistic theory is correct to view this target as a rule-governed system. Given that it is agreed by all parties that the kinds of subsymbolic systems I am interested in are thus not capable of capturing such rule-governed operations, despite their ability to approximate the observable behavior of a rule-governed system over any arbitrary finite span, I now turn to the question of how else they might explain our linguistic capacities.

4 | PUBLIC LANGUAGE: SHIFTING THE BURDEN

Connectionists have thus painted themselves into a corner. On the one hand, it was precisely the ways in which connectionist architectures *deviated* from natural language that made them seem appealing. Defenses of these systems charged classical theorists with claims of over-intellectualizing the mind, of forcing all cognitive tasks into a linguistic straitjacket. Discreteness, categoricity, lawfulness, and other properties of languages did not seem to apply to psychological capacities in general. However, there is at least one psychological capacity which seems indisputably language-like, namely, *language*. In eschewing language-like cognitive architectures, connectionists face deep problems in accounting for humans' most distinctive psychological capacity.

Perhaps the most promising attempt to resolve this problem involves allowing that language is indeed special, an anomalous discrete and rule-governed island in a sea of fuzzy subsymbolic processes, but to relocate this specialness from the mind itself to the environment with which it interacts. Language, on this account, is a *cognitive prosthesis*: by coupling the connectionist mind with the genuinely symbolic system of language, novel capacities are generated. In particular, discrete, categorical representations, can exist as *public, concrete, symbols*, which serve as stimuli for mental operations of pattern-identification, classification, and so on. But we must not infer from these properties of public symbols to features of the mental processes they trigger. The processes are still ultimately of a kind with the subsymbolic, gradable, similarity-based pattern completion characteristic of all connectionist computation. What distinguishes them is that the system they approximate is categorical and rule-governed.

It is important to distinguish two closely related proposals here. One, proposed in various places by Dennett (1989a, 1989b, 1991, 1993), has it that when connectionist systems interact with public languages, the former are re-organized in radical ways, coming to implement a "virtual classical system." This proposal, obviously, is a version of the "connectionist implementation, classical system" approach, and thus is again not targeted by my argument.

The more radically connectionist alternative, developed most extensively by Andy Clark in numerous places (e.g., Clark, 1993, 1996, 1998a, 1998b, 2006; Clark & Karmiloff-Smith, 1993) is that public language can augment our connectionist mind, but not fundamentally alter its operations.

Clark (2006) provides, among many others, the example of training chimpanzees to identify higher-order relational properties from Thompson et al. (1997). These chimps were trained to associate relational properties, such as sameness and difference, with arbitrary symbols. They were shown pairs of identical objects with one symbol, and pairs of contrasting objects with a distinct symbol. Sufficient exposure to these associations, and a suitable schedule of reinforcement, allows them to successfully classify novel stimulus pairs with respect to whether they are identical or nonidentical. Chimps trained in this way were further able to apply this ability to "higher-order" relations, such as the sameness or difference of the relations that objects stand

in. For example, they could classify pairs of pairs of objects according to whether the pairs stand in the same relation. Clark claims that this ability is essentially a result of the introduction of concrete, public symbols to the pattern-completing mind. By attaching a "tag" to a scene, the chimps can attend to features which they normally would not. In normal circumstances, the objects themselves attract all the attention, and so the relations between them are not as salient. But the availability of a repeated symbol which is present only when the two items in the scene are identical forces the chimp to look for more abstract features. In this way, the mind can still essentially function to look for patterns, but public symbols augment it by highlighting, and even creating, different patterns to complete. While vastly more complicated, the hope is that human language can be accounted for in similar sorts of ways.

5 | PUBLIC LANGUAGE MEETS THE POVERTY OF THE STIMULUS

Any proposal of the sort just described brings with it a difficult-to-meet criterion of adequacy. For whatever proposed properties of language it "offloads" to the linguistic environment, there must be a way of identifying these properties *as* properties of some feature of the environment, and (all) human speakers must be sensitive enough to these properties that their behavior can reflect them. There are, however, principled reasons why this will not be possible. In particular, many of the most interesting and central properties of language seem to be constructed, almost de novo, by human minds, absent evidence of these properties in the linguistic environment.

It is important to be clear about the features of language which pose the most severe problems for subsymbolic accounts of mind. Many theorists who discuss the ways in which language can augment our psychological capacities focus on the power of individual words for opening up new possibilities for thought.[4] However, the deepest problem of language acquisition is presented not by symbols, individual linguistic items, but by the linguistic *system*. This makes assessment of such proposals difficult, as it is unclear how they generalize to the more difficult cases of structural aspects of language. But, any theory which purports to save the subsymbolic approach from the problem of language must so generalize. Knowing a language is not merely having a collection of words, but knowing the rules governing their combination. It is these rules, and the system they form, which will generate my argument against subsymbolic theories of mind.

The first worry is that the structural features of language simply do not seem to be properties of any environmental object. A flat-footed, but plausible, ontology might have it that what is found in the environment are waves of compression in the air (in literate societies, we can add to this various kinds of inscription). Of course, the rules of language are not defined over such entities. And, as has long been known (see, e.g., Mattingly & Liberman, 1969, for a review of early results), the mapping between such physical phenomena and linguistic structures is highly complex. One and the same expression can be physically realized in radically different ways and different expressions can be realized in physically identical ways. It seems rather that whatever structure there is in language, be it phonological, syntactic, or whatever, is instead *imposed* on it by human minds. This provides a prima facie case against the idea that the structure of language can be primarily an environmental resource on which the nonsymbolic mind draws.

[4]See for example Clark, 2006; Lupyan, 2012; Lupyan & Bergen, 2016.

However, resourceful opponents of the internalist view of language just proposed (e.g., Devitt, 2006, 2008) have argued that such a flatfooted ontology is misguided. Devitt points out that while the *intrinsic* properties of whatever can be found in our environment are non-linguistic, this in itself is no barrier to locating the relevant linguistic structure outside of the head. One simply has to recognize the relational nature of such properties. Some soundwave can have, in addition to its intrinsic properties, the property of being a word, an auxiliary verb, and so on, in virtue of the relations it stands in to other linguistic expressions, and to the minds of human language users. This proposal, if successful, could thus inhabit the environment with the symbolic structures needed by the strategy under consideration.

For the sake of giving this "offloading" strategy its best shot, I will assume that some story along these lines can be given, and that the environment really does contain linguistic entities with the properties posited by linguistic theories. I think the case can be made successfully against subsymbolic proposals even while granting this. The crucial point is that, in defending a subsymbolic account of the mind, the presence of such linguistic entities is, on its own, insufficient to avoid the problem. It must further be shown that users and learners of language are *sensitive* to such structures. Obviously, not all properties of an item are perceptible. Looking at someone, I can see their size and shape, but not their siblinghood, or their profession. So even if Devitt's proposal about the ontology of language were correct, much more would be needed to show that we can make do without reference to internal symbolic systems in accounting for our linguistic capacities.

As generative linguists have been stressing for over half a century, children's acquisition of language does not merely involve extrapolating from the expressions they encounter. In many cases, the rules of language adopted would seem positively perverse if the only basis for acquisition was the primary linguistic data.

This point could be made with reference to many linguistic phenomena. The example discussed earlier, of the difference between negative sentences with and without semantically motivated auxiliary verbs, provides one case, as our acquired competence here turns on the difference between lexical and auxiliary verbs, which does not seem to have any perceptible signature. Constraints on question-formation provide another famous case.

How, for example, is a child to learn that arguments allow for wh-extraction ("Who will Nasrin kick the ball to?") but adjuncts do not (*"Who will Nasrin kick the ball if throws it?") without the abstract categories of argument and adjunct, which again display no perceptible signs? Beyond syntax, speakers are able to apply phonological rules to sounds they have never encountered in their native languages, on the basis of abstract groupings of phonemes, as when English speaker's recognize that Bach's family could be referred to in the plural as "the Bach*s*," with an unvoiced final consonant (as in "cats"), rather than "the Bach*z*" with a voiced consonant (as in "bugs") or "the Bach*iz*" with an epenthetic vowel (as in "foxes"). Finally, we can see the same thing in philosophically more familiar cases drawn from semantics, such as quantifier scope. That "Two students read every book" is ambiguous, while "Two students read no books" is not, is explained (e.g., by Beghelli & Stowell, 1997) with reference to highly abstract features of the grammatical structure of these sentences and the ways that specific lexical items mandate or preclude certain such structures. In all such cases, it is insufficient to say that these abstract structures or properties are present in the environment. A story must be provided which shows how these properties are identified by the child acquiring the language, and how they are internalized sufficiently well to project to novel environments.

While in these cases, we may, for the sake of argument, grant Devitt's claim that these structures are in some sense found in the environment, it is hopefully clear why this does not do

what the off-loading response to the problem of language requires. Offloading is only useful if the information off-loaded can be reliably acquired and utilized when needed. And there is strong reason to think that will not be the case in these cases, wherein identifying the crucial linguistic properties of these environmental stimuli presupposes precisely the kinds of abstract, repeatable symbols that this maneuver was supposed to bypass. As these properties have no reliable perceptible reflex, it can only be by relying antecedently on the exact classificatory capacities in question that speakers are able to identify these stimuli as auxiliaries, as adjuncts, as voiced, and so on.

The point can be made even more strongly, however, by focusing on cases wherein even on Devitt's ontologically inflationary account, there is simply no environmental correlate of the linguistic properties in question. To take a particularly clear example, consider the findings by Thornton (1991, 1996) and Thornton and Crain (1994) that children growing up in English-speaking environments utilize question constructions not found in adult English, namely medial-wh-questions:

17. Who do you think who is in the box?
18. Who do you think who Cookie Monster likes?

In adult English, these sentences are ungrammatical. Despite this, one third (Thornton & Crain, 1994, p. 218) of the children tested produced sentences akin to (17) and (18). This is not simply a random mistake made by linguistic novices, but is a grammatically reasonable structure, found in languages such as German and Palauan (Chung, 1994), reflecting the pronunciation of the moved question-particle at an intermediate location to which it is moved during the process of raising to sentence-initial position.

For Devitt, the linguistic entities which populate our environments are *conventional* (see his 2006, pp. 181–182). That is, what makes the sentence I am encountering a question, rather than a declarative sentence, is a matter of the public, communicative, conventions of my linguistic community. Whether or not we can make sense of the idea that it is a convention among English speakers that "Bach" should be pluralized with a voiced final consonant, it is clearly *not* a convention of English that "who" should be pronounced in medial position in (17) and (18). Quite the contrary. So, in this case, there is not even a structure in the environment to appeal to in explaining the child's behavior.

The crucial point here, for our purposes, is that the structures generated by the child are not *evidenced* at all in their learning environment. This may be so either because they are present (keeping in line with Devitt's suggestion) but not in a way that a subsymbolic system could extract, or because they are absent entirely, as in the case of medial-wh constructions in English. In the former cases, the examples in question may be conventional, but are not learnable without the positing of internal, symbolic machinery. In the latter, they are unconventional, but are learned nonetheless. This exemplifies the dissociation between the learner's perceptible environment and the structure of their linguistic capacities, which renders the off-loading strategy impotent.

To explain the child's behavior, complex linguistic structures (grammatical hierarchies, with potential landing sites for moved constituents, etc.) must be posited. But the child has no model of anything remotely resembling these structures available to it. So, this structure, essentially detailed as incorporating repeatable items (wh-expressions, structural locations) behaving in rule-governed ways, *must* be contributed by the child themselves.

Simply put, the problem is that, as Clark et al. allow, subsymbolic (e.g., connectionist) systems are fundamentally pattern completers. But language acquisition cannot be modeled as completion of environmental patterns. The environment simply does not provide a suitable perceptible basis to induce from. Rather, the environment provides inputs to specific rule-governed processes. This is a fundamentally different kind of process, and one that requires a classical approach to cognition.

6 | WHY "BOOTSTRAPPING" WILL NOT SOLVE THE PROBLEM

It is occasionally suggested that the problems just raised for subsymbolic approaches to language can be dealt with by appeal to what developmental psychologists call "bootstrapping."[5] Appeal to bootstrapping processes is an empirical hypothesis. It can explain certain kinds of psychological development, in certain sorts of environment, given certain kinds of starting point. That is, bootstrapping is not (despite its etymology) a deus ex machina, capable of getting a learner anywhere from anywhere. To evaluate the claim that bootstrapping processes will enable a subsymbolic learner to behave in ways that approximate those of a symbol-manipulator, we would need to see much more detail on how such a process works than is suggested by the often-vague appeals seen in the literature.

However, the problems for subsymbolic accounts of linguistic competence go beyond these sins of omission. There are positive reasons to think that no bootstrapping model of language will rescue a fully subsymbolic theory of mind. To demonstrate this, I will describe some paradigmatic bootstrapping models from the cognitive sciences; Lila Gleitman's Syntactic Bootstrapping, and Susan Carey's Quinian Bootstrapping. I will then identify some core features of these models, and argue that these are absent from the learning task subsymbolic systems create for language acquisition.

Gleitman et al. (Gleitman, 1990; Gleitman et al., 2005; Landau & Gleitman, 1985; Lidz, 2020; Lidz et al., 2003a, 2003b) have developed a detailed and robustly confirmed theory of how children learn the meanings of novel lexical items. The mere fact that English speakers say "dog," rather than "perro" or "chien," shows that the mapping between word meaning and pronunciation must be learned on the basis of the child's experience. However, this experience seems to massively underdetermine this mapping.

Consider a simple model of language learning according to which a word is learned by identifying correlations between the sound and the presence, or more plausibly salience, of its referent. So, children learn that "chien" refers to dogs because this word is frequently uttered in the presence of dogs, and infrequently uttered in their absence. Aside from the controversial and generally unmotivated empirical assumptions made by this account, it is obvious that it cannot be the general story. Many words simply do not correlate with perceptually available referents. Consider psychological verbs, such as "think," "believe," "hope," and so on. Could there be anything *perceptible* in the child's environment which correlates with the use of these words, let alone which could discriminate between them? Other words may have perceptual correlates, but these correlate with multiple expressions, leaving it a mystery how children are able to associate the perceived phenomenon with one rather than the other. (Almost) every scene

[5]For example, Clark (2001, p. 136).

describable as a *giving* is describable also as a *receiving*, and likewise for *selling/buying*, *chasing/fleeing*, and so on. How then do children do manage to learn these words?

Gleitman argues that this is made possible through *syntactic bootstrapping*, a process by which language learners narrow down the range of descriptions of the scene to associate with these verbs by attending to the sentential contexts in which these words are used. Lidz (2020) gives the example of a child learning the difference between contact verbs ("hit," "kick," ...) and change of state verbs ("break," "melt," ...). Many scenes will be truly described by verbs from both of these classes—a scene in which something is broken is likely one in which something contacts it—which poses a problem for the learner. However, if the child can keep track of the sentential contexts in which such verbs have been encountered, this can serve as a source of discriminating information. Change of state verbs allow for causative alternation, wherein the object of a transitive sentence can serve as subject of an intransitive ("Arturo broke the window" → "the window broke"), whereas contact verbs cannot ("Arturo punched the window" ↛ *"the window punched"). If a child can identify these distributional facts, they can place novel expressions in these classes, despite their perceptual evidence failing to do so. In this way, by identifying the syntactic properties of an expression a child can "bootstrap" their way toward knowledge of the word's meaning, even though mere word–world correlations are on their own insufficient.

Carey (2009) aims to tackle structurally similar problems in the domain of concept acquisition. Analogously to Gleitman's work, Carey aims to resolve the apparent in-principle obstacles to any account of acquisition of concepts on the basis of experience. Fodor (1975, 1981) argued that novel concepts could not be rationally acquired, on the grounds that the only known model of rational response to evidence is a hypothesis-confirmation model, but that such models must presuppose the very concepts that are being acquired. Carey aimed to develop a rational model of concept acquisition which was not based on hypothesis testing.

Even aside from Fodorian worries, there is an obvious conceptual puzzle raised by the acquisition of many kinds of concept, especially examples that Carey focuses on, such as arithmetical concepts. Experience seems in principle unsuited to provide a learner with the information contained in such concepts. Mathematical competence involves recognizing that there is no largest number, but this fact, it seems, simply could not be extracted from any perceptual experience. Similar sorts of puzzle arise with modal concepts, evaluative concepts, and others. Carey developed a theory of "Quinian bootstrapping" as an existence proof that novel concepts could be acquired rationally, despite their underdetermination by perceptual evidence.

Carey's model of numerical concepts works in the following way: A child brings to the table several innate competencies in the ballpark of numerical cognition, crucially for our purposes a subitizing system for immediately, and in parallel, recognizing the number of objects perceptually presented, so long as this number is small (up to around four). The representations output by this system have some desirable properties, as potential building blocks for arithmetic competence, especially discreteness. However, they clearly are insufficient in crucial respects. They allow representation only of very small numbers, and are stimulus-bound. However, children in developed, numerate societies are granted an extra cognitive resource: They are taught the number series. Carey argues that children begin by simply memorizing the series, as they would a nonsense rhyme (*hickory*, *dickory*, *dock*). They then notice correlations between early members of the series and the representations output by the subitizing system: "One" is more frequently uttered when one object is salient. This gradually creates a mapping between the early members of the series and concepts corresponding (partially) to these small numbers. This explains the observation that young children go through developmental stages wherein they

"know" just some small subset of the numbers. "One-knowers" can provide one object when requested, but perform at chance when asked for more than one, "two-knowers" distinguish one, from two, from more-than-two objects, but make no distinctions among the latter category, and so on. However, when they get to around "4" in this mapping process, children near-instantaneously acquire all the remaining concepts corresponding to numbers on the memorized number-line. Quinian bootstrapping purports to explain this final achievement: Children notice relational analogies between the pseudo-numerical concepts applied by the subitizing system and the number series: A representation of two objects stands in a particular relationship to a representation of one object (namely, the relation of *containing one more object*), and "2" stands in a particular relationship to "1" (namely, the relation of *following in the number series*), and likewise for two and three ("2" and "3"), and so on. This parallel motivates an extrapolation: Subsequent numerals in the number line stand for successor numbers. In this way, the child is able to acquire as many concepts for numbers as they have terms in the number series.

With these sketches of fleshed-out bootstrapping acquisition models on the table, we can home in on the distinctive features and presuppositions of such models. First, we can note that these models are aimed at solving exactly the sort of problem that the subsymbolic theorists need to solve with respect to language. Namely, they are designed to explain how novel psychological representations can be acquired despite the perceptual evidence underdetermining the nature of these representations. However, both of the proposals just discussed involve the learner first constructing a representation of the structural/relational properties of the target, which *is* extractable from the learner's perceptual experience. The distributional properties of as-of-yet unlearned verbs (e.g., that they take nominal, but not clausal, complements) is, given that the learner can identify word and phrase types, evident in the primary linguistic data. Likewise, the number line, containing relational information about which numerals follow which, is provided to the child by their teacher. This structure can then be partially filled by semantic information the child already has, such as innately available subitizing representations or knowledge of the meanings of previously acquired expressions. The child then has the easier problem of "solving for" the remaining gaps in the structure.

This creates a significant disanalogy with the task Clark and others have claimed must be involved in language learning. Paradigmatic bootstrapping models exploit the fact that the structural relations and properties needed for the first stage of bootstrapping are perceptually available to the child. In the case of acquiring the structural rules of language, however, the task is not one of attributing semantic properties to the as-yet-uninterpreted parts of partially interpreted structures. Rather, the task is to identify the structure itself. This is a fundamentally different kind of problem. Even if we allow, with Devitt, that public symbols have the often-baroque hierarchical structure attributed to linguistic expressions by linguistic theories, they do so only because they are imposed on them by language users. But this structure itself seems to be suitable only as the input, not the product, of bootstrapping processes.

The major tendency of extended approaches to cognition is to look for ways to offload mental tasks onto the environment. Paradigmatic examples here include utilizing external memory sources, such as a notebook or an iPhone, to reduce the amount of information that must be stored and kept track of in a human brain. What the above case of language shows is that it is crucial, in evaluating such an off-loading hypothesis, to determine which cognitive capacities are needed to extract the relevant information from the environmental resource in question. And in at least this case, the capacities needed to identify the relevant features of the environment are precisely those capacities that such an extended approach was supposed to explain, short-circuiting the proposal.

We can see this exemplified in one of the most developed and sophisticated extended approaches to mind, namely Clark (2015), which weaves together extended, enactive, and predictive accounts of cognition. The core idea is that a mind built to minimize predictive error can use the external world as a supervisor for the development of a generative model of the sources of experience. A perfect such model would, of course, lead to no predictive errors. Learning, for Clark, then involves coming to more and more successfully approximate such a model. Clark calls the development of this model "bootstrap heaven," on the grounds that repeated attempts to create such a model, and compare it to the actually encountered sensory inputs, can allow a kind of hill-climbing capable of acquiring highly complex knowledge of the causal structure of the environment. Clark explicitly applies this model to language (p. 19), claiming that a process aimed at generating accurate predictions about encountered speech is liable to produce a model of grammar.

The problem with this sort of bootstrapping procedure is that it is only by recognizing some stimulus as a piece of language, as a lexical verb, and so on, that we are able to determine whether predictions about it are correct. As we saw earlier, it is these abstract categories that are important for projecting future behavior, not any perceptually available features of the stimulus itself. So, in order to use these stimuli in improving future predictions, a learner must already be able to identify their linguistic properties. This thus exemplifies the general point that for a bootstrapping model to work, the environmental structures which are being leveraged by the learner must themselves be identifiable by the learner. This is not the case in Clark's subsymbolic model.

This lesson may generalize quite substantially to extended approaches to mind. These approaches have been argued to apply to a wide range of areas beyond language. As with bootstrapping models, how successful such models will be depends on the empirical details of the case—specifically, how much can be attributed to the learner at the start of the acquisition process, and whether this suffices for them to "latch on" to the requisite target properties in the environment. Such questions are often overlooked in much of the extended cognition literature. Hopefully the above has shown why facing these questions head on is crucial for the development of an empirically motivated account of mind, and that answering them is often far from trivial.

7 | CONCLUSION

Subsymbolic and extended accounts of the mind have many virtues. And the entreaty to be wary of overly intellectualized or anthropomorphic models of the mind is well-taken. Formal, linguistic, or logical models seem ill-suited for many of the gradable and dynamic capacities of both human and nonhuman minds. However, I hope to have shown that this aversion to symbolic models can overstep its usefulness. Human language is as essentially symbolic as it initially seems. Decades of work in linguistics has established that repeatable symbols and rules governing their distribution are unavoidable if we wish to accurately characterize and explain our linguistic capacities. And given the nature of language acquisition, it seems that no appeal to public symbol systems can avoid this fact. The gulf between perceptible linguistic stimuli and the properties imposed on such stimuli by a linguistic mind is too great to offload the details of the latter onto the former. And the characteristic strategy for acquiring information non-obviously related to perceptible properties, bootstrapping, seems fundamentally unsuited for the task of constructing, as opposed to interpreting, structural representations. So, at least some aspects of the mind still require appeal to classical, symbolic architectures.

ACKNOWLEDGEMENTS

Gabbrielle Johnson, Kevin Lande, and John Dupre all read earlier versions of this article, and provided useful comments. I am also grateful for the extensive feedback provided by two referees for this journal.

DATA AVAILABILITY STATEMENT

There is no data available.

ORCID

Gabe Dupre https://orcid.org/0000-0001-9312-4691

REFERENCES

Beghelli, F., & Stowell, T. (1997). Distributivity and negation: The syntax of each and every. In A. Szabolsci (Ed.), *Ways of scope taking* (pp. 71–107). Springer.

Berent, I. (2013). *The phonological mind*. Cambridge University Press.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, *97*(3), 303–352.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Chomsky, N. (1957/2020). *Syntactic structures*. De Gruyter Mouton.

Chomsky, N. (1966). *Cartesian linguistics: A chapter in the history of rationalist thought*. Cambridge University Press.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(2), 157–205.

Christiansen, M. H., & Chater, N. (2016). *Creating language: Integrating evolution, acquisition, and processing*. MIT Press.

Chung, S. (1994). Wh-agreement and "referentiality" in chamorro. *Linguistic Inquiry*, *25*(1), 1–44.

Clark, A. (1993). *Associative engines: Connectionism, concepts, and representational change*. MIT Press.

Clark, A. (1996). Linguistic anchors in the sea of thought? *Pragmatics & Cognition*, *4*(1), 93–103.

Clark, A. (1998a). *Being there: Putting brain, body, and world together again*. MIT Press.

Clark, A. (1998b). Magic words: How language augments human computation. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 162–183). Cambridge University Press.

Clark, A. (2001). Reasons, robots and the extended mind. *Mind & Language*, *16*(2), 121–145.

Clark, A. (2006). Material symbols. *Philosophical Psychology*, *19*(3), 291–307.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards: A philosophical and psychological perspective on the development of thought. *Mind & Language*, *8*(4), 487–519.

Dennett, D. C. (1989a). *The intentional stance*. MIT press.

Dennett, D. C. (1989b). Mother nature versus the walking encyclopedia. In W. M. Ramsey, S. P. Stich, & D. E. Rumelhart (Eds.), *Philosophy and connectionist theory* (pp. 35–44). Erlbaum.

Dennett, D. C. (1991). Two contrasts: Folk craft versus folk science, and belief versus opinion. In J. D. Greenwood (Ed.), *The future of folk psychology: Intentionality and cognitive science* (pp. 135–148). Cambridge University Press.

Dennett, D. C. (1993). Learning and labeling. *Mind & Language*, *8*(4), 540–547.

Devitt, M. (2006). *Ignorance of language*. Oxford University Press.

Devitt, M. (2008). Explanation and reality in linguistics. *Croatian Journal of Philosophy*, *8*(23), 203–231.

Dretske, F. I. (1977). Laws of nature. *Philosophy of Science*, *44*(2), 248–268.

Fodor, J. A. (1975). *The language of thought*. Harvard University Press.

Fodor, J. A. (1981). The current status of the innateness controversy. In J. Fodor (Ed.), *RePresentations: Philosophical essays on the foundations of cognitive science* (pp. 257–316). MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, *1*(1), 23–64.

Hale, M., & Reiss, C. (2008). *The phonological enterprise*. Oxford University Press.

Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press.

Kam, X.-N. C., Stoyneshka, I., Tornyova, L., Fodor, J. D., & Sakas, W. G. (2008). Bigrams and the richness of the stimulus. *Cognitive Science*, *32*(4), 771–787.

Landau, B., & Gleitman, L. R. (1985). *Language and experience: Evidence from the blind child*. Harvard University Press.

Lidz, J. (2020). Learning, memory, and syntactic bootstrapping: A meditation. *Topics in Cognitive Science*, *12*(1), 78–90.

Lidz, J., Gleitman, H., & Gleitman, L. (2003a). Kidz in the 'hood: Syntactic bootstrapping and the mental lexicon. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a lexicon* (pp. 603–636). MIT Press.

Lidz, J., Gleitman, H., & Gleitman, L. (2003b). Understanding how input matters: Verb learning and the footprint of universal grammar. *Cognition*, *87*(3), 151–178.

Lupyan, G. (2012). What do words do? Toward a theory of language-augmented thought. *Psychology of Learning and Motivation*, *57*, 255–297.

Lupyan, G., & Bergen, B. (2016). How language programs the mind. *Topics in Cognitive Science*, *8*(2), 408–424.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Mattingly, I. G., & Liberman, A. M. (1969). The speech code and the physiology of language. In K. N. Leibovic (Ed.), *Information processing in the nervous system* (pp. 97–117). Springer-Verlag.

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.

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, *8*(1), 1–56.

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, *29*(6), 1007–1028.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & The PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (pp. 216–217). Bradford Books/MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland, D. E. Rumelhart, & The PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (pp. 7–57). Bradford Books/MIT Press.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–23.

Smolensky, P. (1991). Connectionism, constituency, and the language of thought. In B. Loewer & G. Rey (Eds.), *Meaning in mind: Fodor and his critics* (pp. 201–227). Basil Blackwell Ltd.

Thompson, R. K., Oden, D. L., & Boysen, S. T. (1997). Language-naive chimpanzees (*Pan troglodytes*) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes*, *23*(1), 31–43.

Thornton, R. J. (1991). Adventures in long-distance moving: The acquisition of complex wh-questions (Doctoral thesis). University of Connecticut.

Thornton, R. J. (1996). Elicited production. In D. McDaniel, C. McKee, & H. S. Cairns (Eds.), *Methods for assessing children's syntax* (pp. 77–102). MIT Press.

Thornton, R. J., & Crain, S. (1994). Successful cyclic movement. In T. Hoekstra & B. Schwartz (Eds.), *Language acquisition studies in generative grammar* (pp. 215–253). Johns Benjamins.

Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Harvard University Press.

van Fraassen, B. C. (1989). *Laws and Symmetry*. Oxford University Press.

How to cite this article: Dupre, G. (2022). Public language, private language, and subsymbolic theories of mind. *Mind & Language*, 1–19. https://doi.org/10.1111/mila.12400