DELIBERATIO
STUDIES IN CONTEMPORARY PHILOSOPHICAL CHALLENGES

https://deliberatio.uvt.ro/
2025, *5*, 85-100 SSN 2810 – 5826 ISSN – L 2810 – 5532 Print
ISSN 2810 – 5532 ISSN – L 2810 – 5532 Online

Symbols' Individuation: The Challenge and a Possible Solution

Sorin Crețu*

Abstract

This paper investigates the problem of symbols' individuation, emphasizing the major challenge it poses for role-functionalism (Fodor's LOTH) and CTM, while opening toward a possible solution drawn from computational neuroscience. The Frege Cases and the puzzles they generate are salient examples of the difficulty of differentiating between thoughts that "corefer", while still allowing mental states to cause behavior effectively in relation to reality and to preserve relevant, shareable knowledge about the world. The type of content by which symbolic expressions can be individuated may shed light on the possibility of psychological generalizations, but we cannot always admit "tolerable exceptions" in order to preserve the advantages of "multiple realizability". Furthermore, the individuation of neural states by "neuroanatomical features" is supported by the brain's neuroplasticity. In this light, a stronger individuation principle is suggested: "sparsity". It includes symbols' individuation at abstract cognitive levels and is more flexible and robust with respect to generalization (as a neuro-cognitive-informational principle), causality (causal efficacy), and differentiation conditions (between types and tokens). Moreover, "sparsity" genuinely functions in a highly distributed cortical environment driven by neural parallelism. This paper advances this hypothesis, also pointing toward the embodied cognition perspective and its possible extension to AI.

Keywords: symbols' individuation, language of thought hypothesis, computational theory of mind, Frege cases, sparse distributed representations, thousand brains theory

* *PhD in Philosophy, University of Bucharest, visual artist, Laval, Quebec (Canada). E-mail: soleliom@gmail.com. https://www.sorianart.com*

1. Symbols' Individuation – The Challenge

The functional individuation of mental states is based on their functional causal role (Fodor, 1981, pp. 118-119), which stipulates the functional properties that the system is required to have in order to operate and to cause a certain event or result. We recognize a mental state-type that includes a particular mental state-token by the causal role of that mental state-token (Fodor, 1981, p. 118).

For role-functionalism, mental state-types, which are linked by causal relations, are identical to the functional role "itself" (Bennet, 2007, p. 323). Mental state-types are defined by the functional role they fulfill, pointing to functional properties that must meet the specifications of that role. Such functional state-types do not depend on how the system is built, but rather on the functional role they perform in its operation, which allows for the "multiple-realizability" (Fodor) of mental state-types and excludes the identity between mental state-types and physical state-types.

We treat representational mental states as mental symbols, both of which have intentional and propositional content, as well as semantic properties. The causal properties of mental states determine their content and, therefore, the content of the underlying representations. Since mental states are described in terms of their functional causal role, this role stipulates the semantic properties of symbolic mental representations (Fodor, 1981, pp. 122-123). The semantic content of mental representations provides a propositional description of the functional role and is also the carrier of its meaning. As Fodor points out, "a sufficient condition for having semantic properties can be specified in causal terms" (Fodor, 1981, p. 123), which are the terms of the functional causal role.

There are a few important questions to pose. How can we determine the "individuation conditions" of mental states as a function of their functional role or roles? Does knowing the causal role of a mental state suffice to specify the conditions of individuation for that mental state? Is a particular causal role the carrier of the "different individuation conditions"? (Schneider, 2009a, pp. 6-7)

These questions apply to symbols as equivalents of mental states. But what are these symbols, and what roles do they satisfy? To answer this question, we have to clarify what their "individuation conditions" are, as defined by their causal roles. If we can clarify which "features" ("semantic properties" or "neural properties") allow mental symbol-types to be differentiated ("classified"), we can understand the kind of individuation that applies to them.

MOPs (central to LOT), as mental symbols (central to RTM and CTM), are the internal vehicles for our thoughts (Schneider, 2009b, pp. 524, 551), have intentional content, and are the tools through which we can achieve the naturalization of intentionality.

Mental symbols, as mental representations, by virtue of their "semantic features", could be regarded as having meaning; as computational entities, however, they have "computational features" (Schneider, 2009b, p. 526), syntactic in nature.

In order to provide criteria for symbols' individuation, and to know how to "interpret" them, we have to understand which of these features are "essential", so that we can determine under what "conditions" a symbol-token is considered to be included in a certain symbol-type (Schneider, 2009b, p. 527).

Clarifying how symbols individuate is fundamental to understanding how the roles of mental symbols, taken as MOPs (causally efficacious and supportive of the naturalization of intentionality), are fulfilled, so that we can explain thinking and behavior (Schneider, 2009a, pp. 5-6).

One way to individuate symbols is semantically, "by their meaning" (Schneider, 2009a, p. 7), which is held in their content (Fodor). Schneider points to Searle's "Chinese Room thought experiment" as evidence that one can operate with symbols without needing to know their meanings.

If we consider symbol-types to be semantic entities, the prospect of naturalizing intentionality would fail, since that requires "intentionality of thought" to be grounded in a "causal, informational" relation linking the symbol to the external world, which is "physical and non-intentional" (Schneider, 2009a, p. 8).

Taking into account such a semantic condition for symbols' individuation, based on referential content in terms of "broad content" (external individuation of content, as "wide content", along with considering "denotation and truth" as "basic semantic properties of thoughts" (Schneider, 2005, p. 423)), would leave us in a weak position, even if we hoped to specify which "primitive symbol tokens are of the same symbol type", since our goal is to find the physical basis of intentionality and the semantic view would prevent us from achieving it. Moreover, if we try to answer the criterion - does the "same semantic content", making different "symbol tokens", belong to the same "symbol type"? (Schneider, 2009a, p. 7) – and we assign coreferential terms to a symbol token, this will prevent a causally efficacious explanation of different individuals' behaviors (Schneider, 2009a, p. 8), because we cannot guarantee the same knowledge for those individuals.

If we restrict the individuation conditions to the "narrow content", eliminating the referential portion, this will not support the reality of the individual's integration with the environment and would not offer a solution for naturalizing intentionality either (Schneider, 2009a, p. 8).

Semantic individuation conditions therefore pose such difficulties.

A particular view, certainly not orthodox from a functionalist perspective, is "type identity" ("orthographic") individuation (Schneider, 2009a, p. 8), in which tokens of a symbol type might be associated with supposed "particular neural structures" (being tokens of a "brain state type" identical to the symbol type), which

would provide a criterion for how they differentiate as a function of brain states and would also provide a neural mapping of symbols.

However, if we take this path of neural-state individuation, we can take either the route of "neurocomputational roles" or that of "neuroanatomical features," with Schneider referring here to Fodor's position (Schneider, 2009a, p. 9).

For a role-functionalist, symbols' individuation by "neurocomputational roles" of brain states would raise the question of whether it is possible to support cognitive processes with a connectionist argument functioning at the implementation level in the cortex. If this case were proven, it would pose difficulties for arguing that mental states are causally efficacious in generating other mental states, producing an unlimited number of thoughts through a combinatorial structure of symbolic representations—that is, cognitive productivity and systematicity—and explaining behavior (Schneider, 2009a, p. 9).

On the other side, symbols' individuation by "neuroanatomical features", taking into account the neuroplasticity of the cortex, would support a multiple realizability based on brain-state types that instantiate fundamental concepts in duplicate forms in the cortex, particularly as observed in cases of brain damage. It remains to be understood how such particular realizations of symbols (tokens) in brain states (or, going even further, the presumption that such brain states would have "semantic features") can account for their types across multiple human beings, since we all process information differently and react and adapt differently in cases of brain damage or trauma. Here Schneider (2009a, p. 9) refers to Fodor for an individuation based on "neuroanatomical features". This type of individuation would need to account for how we can generalize over symbol-types to make them shareable and communicable.

A possible alternative for these last two variations on type-identity symbols' individuation would be to reassess them in combination with embodied cognition and simulation theories, within a framework less constrained by classical RTM. As long as such approaches remain within a computationalist context, however, they would have limited explanatory power.

The view that defines the core role-functionalist perspective in symbol-types' individuation is centered on their computational role in cognition and on whether such individuation functions across "all generalizations" of a symbol type ("total computational role") or is limited to a portion of the "total role" that a symbol-type has in computational processes (Schneider, 2009a, p. 10).

The former is a powerful tool for the classification of symbol-types, able to reflect an individual's representational structures (MOPs) and to avoid the difficulties of semantic interpretation. Yet total computational-role individuation poses a major drawback. "Psychological generalizations" do not function across different human beings, or even for the same individual "at different times", in such a way that symbols could be "publicly" explained, i.e. the "publicity argument"

(Schneider, 2009a, p. 10). The reason is that we do not share the same symbol-types, since we differ significantly in knowledge, memory capacity, the capability to remember, and "personality traits". Our understanding of a particular symbol-type, and of what it represents in terms of the information it carries from one individual to another, is often localized in groups and populations. Similar behaviors do not necessarily entail that we have the same way of thinking (MOPs) or the same understanding of the symbol-types we individually represent (tokens of them). We might assume that the computational mind is LOT-based, but if this symbolic language cannot be shared, we would not know what it means for us to behave and take actions in the same way or differently. This aspect directly affects our ability to account for the causal mechanisms involved in producing thoughts and behavior, and therefore for their causal efficacy between different human beings (Schneider, 2009a, p. 10).

On the other hand, we have the view of symbol-types' individuation by "limited computational role" (symbols "molecularism"). If the "total computational role" perspective takes individuation over the entire "class" of "computational" causal relations involved in symbol-type generalization, the "limited computational role" reduces that class to those specific instances in which that symbol-type is "tokened", restricting the understanding of that symbol-type to the information ("mental files") shared by a certain group of individuals about that type; in this context, Schneider (2009a, p. 11) stresses that computational relations are "symbol-constitutive". This version of computational-role individuation pushes the account toward "narrow content," as Schneider notes when discussing Fodor's position.

Although such a tendency toward "narrow content" widens the distance between the individual and his/her environment—driving toward the supervenience of mental state-types on "intrinsic states of the individual" (Schneider, 2005, p. 423)—t increases the sharing capability over that tokened symbol-type class and strengthens the individuation conditions. It is a more realistic perspective, since we frequently share and communicate over symbol-types within particularized modes of thinking-MOPs (Schneider, 2009a, p. 11)-and the limited computational relations involved with such a role are causally efficacious and sufficient for that tokening to type the symbol (e.g., specialized types of knowledge, mathematical, art, philosophy concepts, implying refined computational roles over symbol-types, increasingly "fine grained" (tokened), narrowing the mental content, allowing us to "define" and to explain them).

With respect to the objection to this limited computational-role individuation, when attempting to increase the range of the computational relations over a symbol-type, adding more "requirements" for that type and therefore increasing the range of the individuation role, it is true that such an extension would generate different symbol tokenings or "notions" for the same symbol-types for different people (Schneider, 2009a, p. 11). However, what seems, on one side, to prevent

generalizations over symbol-types when expanding the computational relations is, on the other side, perhaps one of the main sources of the creation of art: precisely such extensions, which cannot be generalized, provoke surprise and wonder when they are shared, breaking the expectation that such symbols can be typed and then made objects of prediction in the eyes of observers.

"Strengthening" the requirements to force the typing of the symbol would not work, since we cannot impose the same computational relations on so many different people, devise objective criteria to discriminate certain computational relations against others, or force them on others (unless we plan to force everyone to conceive of the world in the same way). To Schneider's point, we actually end up with different "mental symbols of dog" (Schneider, 2009a, p. 11), which is what actually happens, and therefore with various types of behaviors following such distinct causal computational relations over a representational symbol-type tokened so differently (e.g., if we think, in this respect, of certain traditional Chinese culinary practices as opposed to the Western dog-loving attitude).

If we cannot provide a method of reconciliation between the individuation conditions (tokens) of "identical symbol types", the computational roles over such instantiations of symbol-types will tend to diverge, which will prevent "computational generalizations", since, as demonstrated above, we can provide "counterexamples" (Schneider, 2009a, p. 12) of tokens of the same symbol-type that fall into different sets of computational relations. Because we cannot rule out such exceptions, we will not be able to predict (Schneider, 2009a, p. 12) individuals' behavior in a consistent manner (individuals will behave differently, as they are subsumed under different causal computational relations and symbolic representations) or formulate robust "psychological (intentional) laws". These considerations apply to the "total computational role" as well, with the same consequences, since the limited computational-role view was supposed to be an alternative, perhaps even a solution, to the former one.

Without a theory of symbols' individuation, the entire LOT (which is fundamentally based on "symbolic expressions", on which "systematicity of inference" applies, resulting in the "inferential character of cognitive processes") is at risk and, consequently, so is CTM, as Schneider points out (Schneider, 2009a, p. 12).

2. The Strange Case of Frege Cases

The discussion regarding the conditions of individuation, particularly by broad content, as the externalist view proposes, is important because it reveals yet another major constraint LOT needs to address: the Frege Cases.

The "sensitivity" to the "broad content" of the laws of intentionality (accounting for the structure of the mental states in terms of their principles—for example the

principle of non-contradiction-and defining the mental states in terms of their "directedness" to the world) is founded on the neo-Russellian interpretation that would enable "psychological explanations" to be predictive of individuals' behavior (Schneider, 2005, p. 427). The referents themselves would gain power over how they are referred to, but the situation presented by Frege Cases demonstrates that this power is not fully justified. In order to generalize over such individuation conditions and predict behavior, we need to explain how Frege Cases affect such predictions and resolve the Frege puzzles deriving from the persistence of these cases.

If the Frege Cases are "thought experiments" exemplifying the difference sense ("Sinn") and reference ("Bedeutung") and exploring situations where two expressions bear the same reference but have different senses (meanings), associated with different descriptions and dependent on context (as in the classical case of Planet Venus, referred to by the expressions "morning star" and "evening star"), Frege puzzles point to the noteworthy "cognitive significance" of the difference (both in terms of identity and truth-value) between such expressions or names and the "belief ascriptions" consisting in such co-referring terms (Schneider, 2011, p. 184).

The neo-Russellian (neo-Fregean) view postulates that the propositions (with subject and predicate) of which mental states are made function as the linguistic and logical (in terms of implication, conjunction, and negation) relational medium through which mental states are directed to the world. Such propositions ("Russellian propositions" upholding the strong identity relation between the name and the object referred to in the world) stand in a "binary relation" with the "agent" having that mental state "believing" (Schneider, 2011, p. 183). The implication is significant. Since these LOT expressions, containing co-referring terms (symbols), are, according to the neo-Russellian view, "type-identical", they can be generalized "under the same intentional Laws" (Schneider, citing Fodor).

The problem is that this generalization does not work in the real world when we try to individuate such symbols, because the prediction of behaviors provided through such psychological explanation would fail. Individuals holding such beliefs, in terms of co-referring symbols, will act differently depending on their "way of conceiving of things" and their "knowledge" of such co-reference (Schneider, 2005, p. 423). MOPs, as modes in which intentional agents present the world, are therefore a more realistic alternative for supporting intentional laws and gaining cognitive significance, even if we have to give up the all-powerful desideratum of psychological generalizations. The neo-Russellian propositions are "more coarse-grained than MOPs", and so is the broad content of mental states that they support in order to define the laws of intentionality. We represent the same types of objects in the world through different types of mental symbols in countless

ways, and art and music are certainly among the oldest and most enduring such modes of expression.

On the other hand, the way such co-referring symbols are mentally represented by different individuals and, even more so, the way they materialize through art and music, is far from public and is shareable only to the extent that knowledge of them is shareable. The fine-graining of such modes of symbolic expression goes so deeply into human thought processes that the contents of mental states tend to narrow even further. This is one possible explanation for why many co-referring symbolic representations ("co-referring thoughts") are not communicable (beyond a limited circle of individuals, periods of time, or historical occurrences), are even less generalizable, and why many forms of art and music manifesting them puzzle and surprise us, without our being able to grasp them with full understanding, merely guessing at their meaning and trying to compensate for the missing links in the co-referential relation, based on the knowledge to which we have access in the context of which we are part.

As Schneider points out, without knowledge of the co-referring relation, one would satisfy the antecedent when inferring a behavior from such a co-referring symbolic representation (mental state) but would fail to satisfy the consequent (e.g. If X believes Y and X desires Y, then X will behave in the virtue of Y, unless X behaves in the virtue of Z, without knowing the Y and Z co-refer): "To consider a well-known example of a Frege Case, consider Sophocles' Oedipus, who didn't realize that a woman he wanted to marry, 'Jocasta', happened to be his mother" (Schneider, 2005, p. 424). The neo-Russellian view is even more problematic when the Russellian predicates of such propositions describe properties or relations, instead of just names that co-refer (Schneider, 2005, p. 424).

One solution to this problem is to accept the schism: some laws of intentionality hold, while others fail because such Frege Cases and Puzzles remain unresolved and because the failure of psychological generalizations on account of co-referential symbolic representations is profoundly human. We should be prepared to accept that, no matter how much we would like to advocate for such a psychological generalization and for the full extent of the action of the intentional laws, so that we could build a model to predict human behavior and explain it computationally through the symbolic framework of LOT, we cannot all, as human beings, have the same knowledge of the world. Moreover, perhaps there are other properties of cognition to be taken into account in explaining human behavior, and the computational model is limited to what can be explained through a model such as LOT.

Another solution for a working Russellian psychological explanation model, such that we can include the Frege Cases under the same laws of intentionality and preserve psychological generalizations when individuating the symbol-types (in this instance, individuation of mental states by their "broad content"), is to accept

"ceteris paribus" clauses in the "relevant laws" of intentionality and contain the Frege Cases in these clauses as "tolerable exceptions". We can achieve that either by considering that the Frege Cases are not "genuine counter-examples" to these laws (Schneider, 2005, p. 425, 433)-point 1-or by proving that the consequent of such co-referring relations can be satisfied in a non-intentional manner (for example, computational), when the intentional one fails, such that we can make valid predictions to explain the behavior of the individuals in the respective situations-point 2 (Schneider, 2005, p. 426).

But, if behavior is to be explained from a "broad content" point of view, the solution would be a nuanced approach blending points 1 and 2: Frege cases are "tolerable exceptions", and they are computational in nature.

The "narrow content" would remain explanatory when greater granularity is necessary for the prediction of behavior (Schneider, 2005, p. 426), but by computational role. However, the main challenge with this solution is justifying it.

As Schneider emphasizes, the root cause of Frege Cases resides in the fact that "intentional generalizations" meant to explain behavior are based on both the "broad content" and the "referential properties" of mental states ("direct reference theory"), while disre
garding the "conceptual role" (functional, computational role) of mental states (indicated through MOPs, which are symbolic expressions in LOT) and also the fact, as previously discussed, that type-identical symbols (thoughts) result at the intentional (referential) level in different representational mental states and cause different types of behavior that cannot be explained within the Russellian framework (Schneider, 2005, p. 426). Schneider remarks: "This tension between the representational and the causal is the fundamental problem regarding psychological explanation for a Russellian theory" (Schneider, 2005, p. 426).

While Frege Cases imply the semantic differentiation of co-referring symbols based on the "broad"/ "wide" content of mental states at the "intentional level", LOT enables the non-semantic differentiation of co-referring symbols based on narrow content at the "computational level", focusing on the "computational role" of such mental states (Schneider, 2005, p. 427). If "Broad Psychology" can reunite both levels, intentional (externalist account) and computational (internalist account), then the individuation of such symbols itself can consider both the individuals' environment (which plays a role in defining the content of mental states) and the "internal computations" over symbolic representations, defined by their computational roles (Schneider, 2005, p. 428). In this way, such symbolic expressions as those in Frege Cases would play the same computational (conceptual) role at the internal, narrow computational level (as computational states and "mental processing") and have "different contents" (therefore meanings for the same referent, as linked to different types: e.g. "morning star" and "evening

star") at the external, wide intentional/ referential level, as representational states pointing to "relations to the environment" (Schneider, 2005, p. 428, 429, 440).

However, such a bi-dimensional concept structure (Schneider's perspective), combining the "conceptual role dimension" (in terms of functional, computational role and narrow content) and "referential dimension"-in terms of "broad content" and "mapping (function) from contexts to extensions" (Schneider, 2005, p. 429)-would bring more consistency to psychological explanations, both by accounting for the inner nature of thought and by supporting the "predictive uniformities" required to describe behavior (Schneider, 2005, p. 430, 441).

Furthermore, it is important to understand at which level the "causal powers" reside; we refer here to mental states or events, even if we must remember that they are contingent on the causal powers of the physical states permitting their instantiation, as illustrated by Jaegwon Kim's "overdetermination" arguments (referring to "physical causal closure principle", and the "exclusion argument") and Karen Bennett's "dilemma of reductionism" (Bennett, 2007). The attribution of causal powers is a much broader matter, essential for determining the domain that grants the causal efficacy of mental properties, especially for the non-reductive physicalism (functionalism) that needs to save the idea of multiple realizability of mental state-types from the "physical causal closure of the physical domain's severe counterarguments" (Kim Jaegwon).

However, if we embrace the LOT thesis from the point of view of the four major properties of cognition and pursue, for the moment, the CTM philosophy, we have to understand which dimension/ level ("conceptual role dimension" or "referential dimension") supports the property of cognition that guarantees the causal powers within the functionalist framework; we are referring to the property of "systematicity of inference" of symbolic expressions (representational in nature), which echoes in the "inferential coherence" property of cognitive processes (computational in nature). Mental events/ states are part of causal chains; our thoughts cause other thoughts; the symbolic representations, instantiating mental events through their content, cause other symbolic expressions in sequences, based on their "causal connections" (Schneider, 2005, p. 428) and their combinatorial structure (syntactic and logical). The conceptual level leverages the conceptual role that thoughts play in the "cognitive economy" (a role that is functional, based on "internal computations" (Schneider, 2005, p. 428), and that supports causal powers and efficacy from the computational point of view), and it relies on inferential coherence-"inductive and deductive inferences", up to the sentences expressed (Block, 1986, p. 628)-to ensure causal linkage in the processes of cognition, in complementarity with the referential level, which accounts for the "relations to the environment" and supports causal powers and causal efficacy from the point of view of intentionality (Schneider, 2005, p. 428). Ned Block (whom Schneider also quotes precisely in these respects) points out that the "narrow content" of mental

states at the conceptual-role level would ensure the traceability of computational relations ("internal computations" in LOT) in terms of their causal powers, by actually being the "total causal role" involved (Block, 1986, p. 628), by virtue of the combinatorial nature of symbolic expressions in "reasoning and deliberation" processes, as an interface connecting the "sensory inputs and behavioral outputs" (Schneider, 2005, p. 441; Block, 1986, p. 628). Block's idea behind his theory of meaning (the Conceptual Role Semantics or CRS) is that the conceptual causal role determines the meaning of concepts and the linguistic forms (words or sentences) in which they are expressed, and it is precisely this role that they play in conceptual architecture that is causally and computationally effective.

Symbolic representations are interconnected through inferential systematicity rules and inferential coherence at the level of cognitive processes, and meaning is generated depending on the place occupied in this vast conceptual, representational network. Thus, the referential paradigm of meaning seems to fall outside the scope of CRS; we do not derive meaning through reference, by "broad content". The inferential rules are based on the conceptual roles they play in the conceptual network, which is systemic and computational in nature and based on interlinked cognitive structures. In the light of CRS, the individuation of symbolic, representational structures (which are MOPs as "computational states") by conceptual role (which is a functional and computational role) faces a dilemma. Are these symbolic structures, subject to individuation, to be treated as MOPs (which are broad "computational states that are items in the LOT syntax" (Schneider, 2005, p. 441) and different ways in which we think of, and represent, the same referents to which we refer out there in the world), or are they internal, narrow, inferential "computational states"-as narrow contents, semantic in nature (Schneider, 2005, p. 441)-organized through systematicity and coherence rules, without appeal to the referential portion, as in Block's view? Unavoidably, this dilemma adds stress to the problem of the naturalization of intentionality. The idea is that we cannot talk about intentionality and, therefore, referentiality, without the "thought process" that supports it, which is in turn based on the conceptual role, as illustrated by Schneider's quotation from Block: "So it is the conceptual role of referring expressions that determines what theory of reference is true. [In c]onclusion, the conceptual role factor determines the nature of the referential factor" (Schneider, 2005, p. 442). Schneider counteracts by stating that "such conceptual roles are merely computational states that, only when supplemented with a reference relation linking the internal states to the world, have contents" (Schneider, 2005, p. 442).

MOPs remain fundamental to the laws governing intentionality, because they can direct our thinking in multiple ways toward the world by using our "episodic memory"-recalling past events, contextual, vs "semantic memory" recalling facts, non-contextual, with propositional content, truth-value and inference-based

(Sant'Anna, 2018, pp. 1-3)-and attention. The neural mechanisms studied by neuroscience, underlying intentionality and accounting for the causal efficacy of mental states, are reflected in the bi-dimensional concept structure (Schneider), and the way neural networks activate and function is also a reflection of the Schneider-Block dilemma discussed above. Efforts to naturalize intentionality should take these significant points of view into account.

3. Hawkins's "Thousand Brains Theory of Intelligence"

The "Thousand Brains Theory of Intelligence" (Hawkins, 2021) is a model of "cortical function" proposed by Jeff Hawkins, aiming to provide a comprehensive understanding of how the neocortex, the most evolved part of the "mammalian brain", processes information in such a way that it gives rise to intelligence and consciousness.

Hawkins's theory is set within the framework of computational neuroscience, which attempts experimentally to reverse-engineer the way the brain is structured and functions, and then to create "mathematical models" in order to identify the computational cognitive principles lying at the neocortical foundation and the underlying "biologically constrained learning algorithms" (Hole & Ahmad, 2021, pp. 4, 5).

The theory converges into six main properties of the neocortex, which are then applied toward the creation of "general AI" (AGI), in terms of cortical "data structures" and cortical "architecture".

The first data-structure property of the neocortex examined in the present paper also adopts K. J. Hole and S. Ahmad's account (Hole & Ahmad, 2021, p. 6) of the remaining five properties of the neocortex-two data-structure-related properties ("realistic neuron model", neocortical "reference frames"), and three neocortical-architecture-related properties ("continuous online learning", "sensorimotor integration" and "single general-purpose algorithm" or "common cortical algorithm")[2] - is "sparsity", or the capability of generating "sparse data representations" (SDRs), to "represent" and "process" information throughout cortical regions. In order for the brain to function efficiently and to reduce the amount of neural activity while obtaining maximum output in real time under changing conditions, it has to use computational resources economically (Ferrier, 2014, p. 4). These representations are the means of realizing this desideratum. "An SDR is a set of binary vectors where a small percentage of 1s represent active neurons, and the 0s represent inactive neurons." (Hole & Ahmad, 2021, pp. 7, 12). The "sparse activity patterns" involve a low percentage of neurons being "active"-

[2] These properties have also been discussed in Chapter 4, "The Representation Conundrum" of my PhD thesis (Crețu, 2024).

having "1" data values-while the rest are maintained in an "inactive" state-with "0" data value (Hole & Ahmad, 2021, p. 7). This neural activity of such a "small fraction of the cell population" within a cortical region "acts as representation" (Ferrier, 2014, p. 4). Sparse representations entail that, at any moment in the neocortex's information processing, "only a small percentage of values are non-zero" (Hole & Ahmad, 2021, p. 5)—as opposed to dense representations, which maximize this percentage (Ferrier, 2014, p. 4)—"up to 50% of cells within an area act as a representation", as in deep-learning artificial networks. In a sparse representation, most of the values in the data vector are zero, and only a few non-zero values are used to capture the important information. Such sparse representations, as the "primary data structure used in the neocortex" (Hole & Ahmad, 2021, p. 7), enable it to gather essential data rapidly and to compensate for internal computational "errors" and for "noisy data from the environment" (Hole & Ahmad, 2021, p. 6). Sparse representations are more "robust" to data noise than dense representations (Hole & Ahmad, 2021, p. 7), as non-zero values are less likely to be affected by random data variations. While dense representations may privilege data accuracy and availability, and may offer a better description of more complex relationships and detailed distinctions between data characteristics than sparse representations, the latter represent data more effectively than the former, because they require less storage space and fewer computational resources. Sparse representations are also characterized by a high degree of interpretability, featuring the most significant properties of the data and providing visibility into its underlying structure. While the "bit positions" in dense representations do not have "semantic meaning" (Hole & Ahmad, 2021, p. 7), in sparse representations they do, signifying a specific property of the object represented.

The cortical sparsity feature is highly significant from the point of view of how represented information is differentiated and classified based on its particular properties, as "dynamic set(s) of patterns" are progressively distinguished from "noisy data" (Hole & Ahmad, 2021, p. 7).

4. The Sparsity Principle and SDRs – a Possible Solution

Sparsity may offer a promising candidate for clarifying how symbolic representations used by the neocortex at high abstract levels are individuated in the modeling of the world. As we can recall, finding the solution to symbols' individuation was one of the major challenges Fodor's LOT encountered on functionalist grounds. To reiterate, sparsity is the first "data structure" property of the neocortex, directly related to how information is represented and classified. In CTM-LOT, mental representations with symbolic structure (LOT symbols) carried the content of a certain mental state (e.g. believing), while being "locked onto" the properties of the objects represented. The symbols' individuation by

"neuroanatomical features" mentioned earlier in this paper was looking for such a feature, and "sparsity" fills this role.

Moreover, "sparsity" could be the most important property of cognition deeply rooted in biological plausibility, competing with the four-property set proposed by Fodor and Pylyshyn within the computational framework ("productivity", "systematicity", "compositionality" and "inferential coherence") that supports LOTH and that, possibly, could supervene on the principle of "sparsity".

Furthermore, given the direct inclusion of "sparsity" in the broader property of such representations as distributed, and given the fact that distributed representations are "combinatorial" in nature, as well as enabling generalizations across representational structures "based on the degree of overlap between their respective patterns of activity" (Ferrier, 2014, p. 5), "sparsity" acquires an even more individuative character.

While, in the "localist representation" view, "the activity of a single cell corresponds to a given percept or concept" (Ferrier, 2014, p. 4), the representation of information uses distinct values (part of different categories or classes) in the form of discrete representations representing categorical information. In the "distributed representation" view, representations differentiate (individuate) "progressively by gradual weight modifications", with a high degree of "redundancy", which enables them to be "fault tolerant" (Ferrier, 2014, p. 5).

In discrete representations, which are more suitable for language processing (words can be thought of as belonging to distinct categories based on their grammatical function or meaning), data can take on only a finite set of values. This aspect is aligned with Fodor's theory, in which cognitive performance is based on the productivity-of-thought property, where a limited number of "atomic" representations can generate an infinite number of complex "molecular" representations. This cognitive capability is grounded in the linguistic nature of thought (LOTH), which underlies the "combinatorial structure" of mental representations and the systematic symbolic nature of such molecular representations, viewed "as symbol systems" ("discrete inner states"). These stand at the foundation of cognitive processes, based on algorithms made of "explicit" rules that "manipulate" such arrays of symbols. The LOT perspective is to be included in this "localist representation" view, which is not representative of how the neocortex functions globally.

This points to a serious reason why psychological generalization over symbolic representations in LOT poses difficulties. The individuation of symbols must function over all generalizations across symbol-types in order to support the "total computational role" criterion for symbols' individuation, which is the ultimate goal of the role-functionalist rationale: to explain computational processes broadly. The criterion of "computational role" is therefore "limited", at best, in successfully individuating over symbol-types.

The computationalist model, which describes the mind as a computational system operating over symbolic representations and cognitive processes as the processing of strings of symbols, with information encoded in such symbolic expressions in a "localist" manner (the concepts of which our knowledge is made are represented as distinct, discrete symbols), faces the connectionist model, which is built on an architecture centered on parallel distributed processing (PDP). In the connectionist model, cognition results from the interaction of multiple interconnected neural networks that process information in parallel, and knowledge is distributed and represented by patterns of activation across various neurons (nodes), of which only a small fraction are activated at a certain moment ("Sparsity"). Connectionist models such as deep-learning neural networks often employ distributed sparse representations.

Arguments based on the distributed, sparse character of the representations produced by the neocortex are gaining traction in explaining cognitive processes and behavior. The brain recognizes patterns and makes predictions based on continuous streams of information received from sensory inputs, data that are encoded and represented with continuous values and processed, in an unsupervised manner, across a large number of neurons working together in parallel. These data are then discretized into sets of discrete values when required to represent high-level categorical information, in a serial, sequential manner, at the level at which we recognize symbolic types, content-based. In terms of Dennett's "multiple drafts" theory (Dennet, 1991), this is the level of "deliberative" thought in the consciousness workspace and, in terms of Hawkins's "Thousand Brains Theory of Intelligence", this would be the post-voting effect or outcome when a minimizing uncertainty "consensus" has been reached over various perceptual or conceptual models.

SDRs are dedicated data-classification and predictive cortical tools ("multiple simultaneous predictions"), capable of learning "temporal sequences" (Hole & Ahmad, 2021, p. 7), and are used in the "Hierarchical Temporal Memory" (HTM) model used in AI, as a "specific realization" of "Thousand Brains" Theory (Hole & Ahmad, 2021, p. 6).

There is a strong correlation between sparse representations and distributed neural networks in the neocortex in terms of the optimization of information processing. In the brain, information is represented using sparse codes, sparse coding being a mechanism designed to process and store information efficiently. In distributed neural networks, information is processed and represented in parallel across a large number of interconnected nodes or neurons, being integrated across multiple levels. Each node in the network is responsible for processing a subset of the input data, and the output of each node is combined to produce the final outcome.

"Sparsity", allowing for a distributed, differentiating character of representations, may better capture symbolic thought and data classification, providing a more flexible neural-based framework. This approach sees cognitive processes as the outcome of various interconnected neural networks processing information in parallel, ensuring "redundancy", "fault tolerance", and more subtle generalizations at the symbolic level. "Sparsity" becomes a new cognitive baseline, possibly guiding mind-brain supervenience toward a renewed perspective.

References

Bennett, K. (2007). Mental causation. *Philosophy Compass, 2*(2), 316–337. https://doi.org/10.1111/j.1747-9991.2007.00063.x

Block, N. (1986). Advertisement for semantics for psychology. *Midwest Studies in Philosophy, 10*, 615–678. https://doi.org/10.1111/j.1475-4975.1987.tb00558.x

Crețu, D. S. (2024). *Integrating art, neuroscience and artificial intelligence* [Doctoral dissertation, University of Bucharest]. University of Bucharest Doctoral Studies. https://doctorat.unibuc.ro/events/cretu-dumitru-sorin/

Dennett, D. C. (1991). *Consciousness explained*. Little, Brown and Company.

Ferrier, M. R. (2014). *Toward a universal cortical algorithm: Examining hierarchical temporal memory in light of frontal cortical function*. arXiv. https://doi.org/10.48550/arXiv.1411.4702

Fodor, J. A. (1981). The mind-body problem. *Scientific American, 244*(1), 114–123. https://doi.org/10.1038/scientificamerican0181-114

Hawkins, J. (2021). *A thousand brains: A new theory of intelligence*. Basic Books.

Hole, K. J., & Ahmad, S. (2021). A thousand brains: Towards biologically constrained AI. *SN Applied Sciences, 3*(8), Article 743. https://doi.org/10.1007/s42452-021-04715-0

Sant'Anna, A. (2018). Episodic memory as a propositional attitude: A critical perspective. *Frontiers in Psychology, 9*, Article 1220. https://doi.org/10.3389/fpsyg.2018.01220

Schneider, S. (2005). Direct reference, psychological explanation and Frege cases. *Mind & Language, 20*(4), 423–447. https://doi.org/ 10.1111/j.0268-1064.2005.00294.x**.**

Schneider, S. (2009a). LOT, CTM and the elephant in the room. *Synthese, 170*(2), 235–250. https://doi.org/10.1007/s11229-009-9581-1

Schneider, S. (2009b). The nature of symbols in the language of thought. *Mind & Language, 24*(5), 523–553. https://doi.org/10.1111/j.1468-0017.2009.01373.x

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