JNphi

How the Physical Symbols Systems Hypothesis and the Modularity Hypothesis are Reformulations of Each Other

Sandro Skansi

Abstract

This paper examines the relationship between two influential theories in cognitive science and philosophy of mind: Fodor's modularity of mind hypothesis and the Physical Symbol System Hypothesis (PSSH) proposed by Newell and Simon. Although modularity is traditionally framed as a theory of cognitive architecture and the PSSH as a computational account of intelligence, this paper argues that the distinction between them is largely conceptual rather than substantive. The central claim is that modularity and the PSSH are, in effect, reformulations of the same underlying theoretical commitment. The analysis begins with a critical review of Fodor's defining criteria for modules, such as domain specificity, mandatory operation, informational encapsulation, and shallow outputs. While these conditions were intended to provide an empirically grounded account of mental organization, many of them remain vague or metaphorical. Nevertheless, taken together, they imply that cognitive processes must be decomposable into functionally distinct components with standardized input–output relations. Such decomposability presupposes the manipulation of structured representations. From this, the paper derives its first key claim: any modular cognitive process necessarily involves symbolic processing. Modules can only function if their outputs are formatted in a way that allows systematic recombination and communication with other components. This requirement aligns directly with the notion of a physical symbol system. Modularity, therefore, implicitly assumes the symbolic ontology articulated explicitly by the PSSH. This establishes a bidirectional equivalence: modularity entails symbolic processing, and symbolic processing entails modular realizability. The paper concludes that long-standing debates opposing modularity to symbolic approaches rest on a false dichotomy, and that recognizing their equivalence offers a clearer framework for understanding both human and artificial intelligence.

Key Words: modularity; artificial intelligence; physical symbols system hypothesis, encapsulation, isotropy

DOI: 10.5281/zenodo.20052107

Corresponding author: Sandro Skansi
Address: Dept. of Philosophy and Culture, Faculty of Croatian Studies, University of Zagreb
e-mail ✉ sskansi@fhs.hr

1. Introduction

There are two seemingly unconnected ideas in cognitive science, one is the modularity theory by Fodor (1983) and the other is the physical symbols system hypothesis or PSSH (Newell and Simon, 1976). Both have proven to be highly influential, but to the best of our knowledge, no one has analyzed how these ideas overlap, and this is our goal in this paper. In a certain sense, this natural idea today has a concrete realization--agentic AI systems, but we will not explore this further, and instead we point the interested reader to (Šekrst, 2025). Modularity itself is highly debated in recent philosophical research (Egeland, 2024; Clarke and Beck, 2023), and research on modularity and cognition is also prolific to this day (Pietraszewski and Wertz, 2022). An interesting discussion, which can be viewed as a "lemma" between modularity and the PSSH, is the idea of module encapsulation (Brooke-Wilson, 2024).

2. Domains, symbols and modules

Even though the idea of the modularity of the mind was always explored as an idea, it was Fodor who put forth an extensive definition (Fodor, 1983: 37-39). Even though he presents a list of these requirements, he softens them a bit by noting that most of them should be fulfilled. Fodor's requirements for modules are: (1) domain specificity, (2) mandatory operation, (3) limited central accessibility, (4) fast processing, (5) informational encapsulation, (6) shallow outputs, (7) fixed neural architecture, (8) characteristic and specific breakdown patterns, and (9) characteristic ontogenetic pace and sequencing.

Condition (1) can be viewed as a spontaneity in accepting only specific kinds of input, like for example the cheater detection module (Bermudez, 2020) not processing inputs such as music. This filtering for modules is automatic, and the keen reader might already see symbolic processing happening in the very idea of modules. But there is more here than meets the eye. When considering a module, the very fact that it is selective about the features it processes suggests that the line between e.g. perception and processing might not be as clear as we thought. Even though there was significant research on this demarcation, like (Clarke and Beck, 2023), the stark distinction between seeing an image (perception) and seeing an image (mind) might exist only in the description of the neurological process, and not really be present in the neurons themselves. By extension, the compartmentalization of the mind into modules might be nothing more than the compartmentalization of the description itself.

This is closely connected to condition (2) as well, since, as Fodor (1983: 52-3) points out, the processing of module inputs (once selected) is not optional. This means that the module for processing speech, processes all speech it receives. Even though we know from experience that we

are able to ignore a given voice which is speaking to us, what Fodor suggests is slightly different. The speech processing module automatically processes speech, unless another model (e.g. an attention module) does not actively ask the initial module to disregard the results of its processing. The processing itself will happen--we will hear and process that voice by default, and only by actively refocusing on something else, we can ignore it. Mandatory processing enables simplicity in modules, as argued in Brooke-Wilson's (2023), but it also enables the idea of modularity itself (Šekrst, 2025). As seen in (Vaswani et al., 2017), the basic idea of attention is computationally equivalent to memory, forgetting, and reformatting combined. Even though what is discussed there is just a computational model, it still holds true that, computationally speaking, basic attention is a very powerful mechanism, capable of replacing a number of features. Having a selectivity like attention might render the whole idea of modules superfluous when considering cognition as a computational phenomenon.

3. Encapsulation and shallow inputs

The condition (3) (Fodor, 1983: 55) is the one that stipulates "a limited central access to the mental representations that input systems compute", which is by itself clear enough, but was abused over the decades. Fodor's idea was not that there is no central processing, but that central processing does not "micromanage" modules, and does not have access to their internal representations. That being said, the very idea of a central processing seems wrong, since it provides a limited explanation by referring to modules, while leaving the more complex processing to the "central" processing, which again is not explained in adequate details. The problem is that with such an approach, one wonders what would justify the very introduction of modules. In a sense, Fodor's module theory is interesting only if the totality of the mind could (one day) be explained solely by modules. But even then, there is a dark cloud looming on the horizon: such modules would have to be connected by a wiring diagram, and that diagram itself is a symbolic system, which processes information, and consequently falls under the PSSH.

Šekrst (2026) notes that there is an interesting connection between conditions (3) and (5), the condition (5) stipulating that input systems are informationally encapsulated (Fodor, 1983: 64-65). Šekrst here is highly benevolent towards Fodor's complete lack of formal and mathematical rigor, since condition (5) is obviously implied by any nontrivial reading of condition (3). Fodor's lack of rigor takes us back to condition (4), which stipulates that input systems are [sic!] fast (Fodor, 1983: 61). Fodor's discussion of this condition is almost comical, by noting what he calls "a paradox". Fodor finds paradoxical that we can spend hours on understanding a philosophical paper, and yet its content is simpler than the cognitive task behind a saccadic eye

movement. A cynic would argue that this would surely be a paradox if the paper was Fodor's, but this would miss the point as much as Fodor's argument did. Fodor seems to ignore the fact that when talking about the cognitive aspects of a saccade, i.e. the complexity presents there, we have got almost to the neural level. In contrast to that, when we are discussing the cognitive processing involved in reading a philosophical paper, even in the most optimistic of cases, we have barely decomposed the ideas in the paper into a flowchart. Therefore, comparing complexities makes little to no sense. This complements well the criticism of the modularity hypothesis put forth by (Pietraszewski and Wertz, 2022).

Condition (6) stipulates that modules have "shallow" outputs. Even though no effort has been made by Fodor to clarify the meaning (Fodor, 1983: 86), there might be an intuitive idea of what "shallow" means here. Consider the idea of a power socket, and call that a "shallow" output of the electric network. By contrast, if there were no power sockets, the only conceivable way would be to have a line straight from the power plant right down to the lightbulb. A lightbulb connected to a socket has a shallow inlet to the system while a lightbulb with a wire going straight to the source has a deep inlet (input). As one module's output is another module's input, the distinction does not matter much. In a sense this condition seems to be crucial for modules, in the sense that the output/input has to be formatted appropriately, supposedly more like a JSON than like the result of a convolutional layer in a neural network. The very structure stipulated here is again symbolic in nature.

4. Neural architectures and learning

The seventh condition (Fodor, 1983: 99) tries to pinpoint modules in specific neural architecture. Of course, this is still more or less wishful thinking, but it seems to reek of symbolicism. To see this, consider the opposite: what would be the connectionist best case scenario? Not now, but in a millennium of research. It would be plausible to conclude, based on the basic idea of connectionism (Šekrst, 2025), that the connectionist ideal is one neuron which is multiplied and connected a large number of times between its copies. This would form the architecture, which should carry the least importance. The main thing for connectionism would be the learning of weights. This is the connectionist ideal. The learning itself should be based on learning ideal weights, w^*. It is worthwhile noting that this conception of the mind suffers from a weird consequence, which seems not to be the case for the human mind in general, although some versions of it might seem oddly familiar. If the initial setup is xw_0, where x is the input and w_0 are the initial weights, and the ultimate goal is to learn some ideal weights w^*, such that wx^* gives the desired behaviour, then a weird consequence follows. In fact, xw^* can be rewritten as xaw_0, i.e. $w^*=aw_0$ for some scaling factor a. But then, if $xw^*=c$, and $xw^*=xaw_0$,

it follows that there is a transformation of the inputs that requires no learning at all for the same result, that is $x'w_0=c$. Moreover, we can actually specify x' as xa, which in turn means that if the connectionist view of the mind is correct, there is no essential difference in learning and simply applying a transformation on the input. Even though in full this sounds highly improbable, there are events of this happening, such as students taking note while reading a textbook and being able to learn much quicker the same content. A second example of this kind of processing would be how children prefer cartoons to reality, due to higher contrast and sharper lines. It is interesting to note that putting all the eggs in the basket of learning is not a uniquely connectionist approach (although this is an essential part for any connectionist theory), as e.g. Prinz (2006) tries to reconcile modularity with (statistical) learning.

The eight condition seems to be the most noteworthy, as it is quite unique and easily testable. Fodor claims that modules exhibit specific and characteristic breakdown patterns (Fodor, 1983: 99). This is interesting, since for the first time Fodor offers something which is empirically testable, and moreover which intrinsically connects modules of the mind with neural tissue of the brain. In a sense, Fodor assumes a mapping between the brain and the mind, albeit this mapping need not be a one to one. The symbolic interpretation here is trivial, since the breakdowns have to be specific in the symbolic process chain.

Condition (9) sounds almost like a bad philosophical limerick: *the ontogeny of input systems exhibits a characteristic pace and sequencing* (Fodor, 1983: 100). First, we have "ontogeny" which should have been "ontogenesis", i.e. the evolution of a metaphysical entity. An interested reader would be puzzled by this term, and for good reason. The very idea of introducing an "evolution" while still describing modules is wild-- comparing it to running before learning to walk would be a collocational understatement. Next comes the "characteristic pace and sequencing". Mind you, we are still describing what modules are! Referring to anything "characteristic" of them in their very definition is a circularity under even the most benevolent of interpretations. How could we understand something "characteristic" of an entity even before the entity is defined? If the "characteristic" is defined by the entity (e.g, "the characteristic quack of a mallard"), then the entity itself cannot be defined in terms of this: we cannot define a mallard by its characteristic quack, since the mallard defines the (characteristic) quack. Next comes the "pace". What on Earth would be "the pace" of a cognitive system? Surely not the "speed" which was already used and abused in a previous condition--but "pace". It defies belief that anyone could use the term "pace" to even tentatively describe a process which happens in microseconds and is also--by their own account--automatic! (which invalidates the second intended meaning of "pace", viz to time one's actions). Lastly, we have "sequencing", whatever that

might be in a system which is interconnected in multiple nontrivial ways and certainly not sequential.

Almost as if he is mocking the reader and testing their intelligence, Fodor exclaims that modularity is “the general fact about the organization of the mind” (Fodor, 1983: 101), only to continue to establish that there should be a non-modular central processing system. According to Fodor this is because modules lack the capacity to process Quinean holism and isotropy. The reader might ask themselves why would modules have to be able to explain Quinean holism, which is *de iure* an epistemic theory, and *de facto* a metaphysical theory about the nature of knowledge, while Quine himself showed that metaphysical theories are empirically underdetermined (Quine, 1975) is again a puzzle for anyone who tried to read Fodor, and yet to this mystery we will never get an unequivocal answer. In any case, Fodor does not seem to have any doubts whatsoever on the nature of knowledge.

5. Isotropy and modularity

Unlike "Quinean holism", isotropy (the idea that any information might be important) warrants a closer look. Fodor tends to see this as an important piece of evidence for a central processing, and the argumentation seems quite clear: if modules are specialized, there needs to be a central, broad, processing which can, at the very least, mark the important parts of information so that modules might process it. The argument seems plausible at first, but it can be quite easily seen that this is just a selection bias in disguise: even though (in theory) we might know out of all information we gathered what parts are important, there is literally no way of knowing whether the information we have not gathered is important. Only under a very dubious assumption that all problems are solvable might we conclude from our inability to solve a problem that we are missing important information. But if we do not make this assumption, as we shouldn't, the inability to solve a problem does not give away any clues about whether we have all the important information or not. But all of this is highly optimistic. The idea that isotropy is real, and as such provides evidence to the need for a central processing, is misguided. It rests on the false assumption that people can extrapolate important information and only after that form a proposition or conduct an action. More often than not, propositions are formed beforehand and only after their formation, the mind tries to find "important" information on which to anchor them. The basis of belief formation is first and foremost habit, and second learning from mistakes. If no mistake is made, the belief is formed purely on habit and not based on "important" information. If there is no important information to be processed prior to forming the proposition, there is no need for a central processing to search for it. There is also a minor point here, which goes hand in hand with our analysis. The "importance" of the

information is relative, and dependent on the proposition or action. In this way, epistemically, the proposition itself precedes the marking of any information as "important".

One could be forgiven for thinking that the discussion became more focused in more recent times, but there are still authors who parrot away the Fodorean doctrine. One prime example is (Robbins and Drayson, 2025). They seem to find the following line of argumentation perfectly acceptable:

1. Central systems handle belief fixation

2. Belief fixation is both isotropic and Quinean

3. Such information cannot be informationally encapsulated

Therefore:

4. Central systems cannot be modular

This argument has a decent chance of going down in history as a lesson of what is bad philosophical argumentation. Let us start with the conclusion. There is little or no sense in calling central processing modular or not, since there exist theories which consider massive modularity with no central processing, and modularity with a central processing. The former has no central processing, while the latter has a central processing alongside modules. No one ever claimed that the central processing is modular. Literally no one. So, in this sense, such a conclusion, which indeed follows from the premises, is nothing new. One should notice how these authors framed their thinking as a logical inference to make the whole argument look more formal and more convincing.

As for the premises, it is really hard to find faulty argumentation where all the premises are false, but Robbins and Drayson sure managed to find this gem. Take premise (1). One might wonder just why you would need central processing and not a module to handle X, and in this case X is "belief fixation". A more general argument would, along the lines of (Skansi and Šekrst, 2021), that any process that could be precisely described with phases (or segments), can be broken into those phases (segments). Modules are an excellent application for this idea of compositionality, since they are by definition idealized components of a process.

Let us turn to premise (2). Belief fixation indeed is "Quinean", *if* Quine's holism is the scientifically confirmed theory of belief formation. But then again, belief fixation would be "Tomean", "Dickean" or "Harryan", for any Tom, Dick and Harry whose theory of belief formation would turn out to be (empirically) the correct theory. As for isotropic, one may only wonder why the general idea that any information might be important would be directly relevant for a mental process called belief fixation. Mind you, not some important information, but the (meta)idea that some information is important.

Continuing to (3) we have the dogma "such information cannot be informationally encapsulated. If one returns again to the idea of isotropism, which says that some information is important (and by extension that some is not), one would be hard pressed to find a better example of informational encapsulation--than a mental faculty collecting important information, or even better, a mental faculty collecting information, some of which is by its nature "important". We are not the first to identify these problems. Pietraszewski and Wertz (2022) had the same skepticism, although they went a step further and identified a number of category mistakes, relating to confusing different levels of analysis. Although as Egeland (2024) rightfully points out, their arguments tend to be quite schematic, and seems to fall short for avoiding future problems, we feel it is a step in the right direction, and that more research critical of Fodor and his acolytes is needed to identify and hopefully fix the argumentation mistakes disseminated by Fodor and his minions. So, how is modularity of this kind a reformulation of the physical symbol system hypothesis (PSSH)?

6. From massive modularity to the PSSH and back again

As we promised, we will show how the modularity hypothesis is just a reformulation of the physical symbols system hypothesis put forth by (Newell and Simon, 1976: 116): (PSSH) "A physical symbol system has the necessary and sufficient means for general intelligent action."

It might be worthwhile to explain what the symbols mentioned in the PSSH could be. Even though anything can be a symbol, like 0 or 1 or letters, or even gestures, in these cases some would be hard pressed to argue that the PSSH does not hold. The interesting cases are the complex symbols like "barking dog" or "a sarcastic comment", which are composed of simpler symbols. This compositionality is the reason why our argument works. Returning to the PSSH, it is important to note what was meant with its formulation. The sufficiency clause claims that a physical symbol system is all that is needed to produce (general) intelligence. But for our case, the more interesting part is the necessity claim. It says that (human) general intelligence is in fact nothing more than symbol manipulation. This leads us directly to the following:

(A) Any modular process is symbol manipulation

As we have established before, and as Fodor argued, modules only make sense if the underlying process can be decomposed. If it can be decomposed, it is composed of parts, which have a structure like Lego bricks, in the sense they fit one another (this is what Fodor called "shallow inputs" and "encapsulation"). We now turn to the other direction. Since modularity aims to explain general intelligent behaviour, it can be stated:

(B) Any general intelligent process can be realized as a modular process

By combining the equivalence stated in the PSSH with implications (A) and (B) it immediately follows that modularity is the same as PSSH. Due to the limitations of the English language, the formulation which itself contains the PSSH might seem peculiar and clumsy, but the equivalence can be easily checked for validity based on (A) and (B).

7. Conclusion

In this paper we have explored the ideas of modularity of the mind, and how they relate to the earlier idea of Newell and Simon, the physical symbols system hypothesis (PSSH). As we have shown, modules were not well defined initially, but several authors contributed and amended the initial theory. Despite their efforts, modularity still has problems, and in the most benevolent of interpretations, modularity is just a nontrivial preformulation of the PSSH. It is easy to see that from the very idea of modularity it follows that any intelligent process can be realized as a modular process, and that the PSSH provides an equivalence between symbol manipulation and general intelligence. What is left, and this was our main contribution in this paper, was to show any modular process is symbol manipulation, and this was shown by carefully analyzing the definition of modularity. From this, the equivalence between the PSSH and modularity immediately follows.

Funding

This research was partially supported by the grant LIVEC (Life-long Development of Emotional Competencies) and partially by the grant AI-COM (Umjetna inteligencija: novi sugovornik hrvatskog drustva)

References

Bermudez JL. Cognitive Science: An Introduction to the Science of the Mind (Third Ed.). Cambridge University Press, 2020.

Brooke-Wilson T. How is perception tractable? The Philosophical Review 2023;132(2): 239–292.

Clarke S, Beck J. Border disputes: Recent debates along the perception–cognition border. Philosophy Compass 2023;18: e12936.

Egeland J. Making sense of the modularity debate. New Ideas in Psychology 2024;75: 101108.

Fodor, JA. The Modularity of Mind. MIT Press, 1983.

Newell A, Simon HA. Computer Science as Empirical Inquiry: Symbols and Search. Communications of the ACM, 1976;19(3): 113–126.

Pietraszewski D, Wertz, AE. Why evolutionary psychology should abandon modularity. Perspectives on Psychological Science 2022;17:465–490.

Prinz JJ. Is the mind really modular? In: Stainton R, ed. Contemporary Debates in Cognitive Science. Blackwell, 2006:22–36.

Quine WV. On Empirically Equivalent Systems of the World. Erkenntnis 1975;9: 313–328.

Robbins P, Drayson Z. Modularity of mind. In: Zalta EN, Nodelman U, eds., The Stanford Encyclopedia of Philosophy (Fall 2025 Edition). Metaphysics Research Lab, Stanford University 2025. https://plato.stanford.edu/archives/fall2025/entries/modularity-mind/

Skansi S, Šekrst, K. The Role of Process Ontology in Cybernetics. Synthesis Philosophica 2021; 36(2): 461-469.

Šekrst K. The Illusion Engine: The Quest for Machine Consciousness. Springer, 2025.

Šekrst K. Fodor's Modularity and Agentic AI: Cognitive Architecture Meets Computational Reality. Unpublished, 2026.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems, 2017;30:5998–6008.