Beyond Interpretability and Explainability:

Systematic AI and the Function of Systematizing Thought

MATTHIEU QUELOZ

Abstract: Recent debates over artificial intelligence have focused on its perceived lack of interpretability and explainability. I argue that these notions fail to capture an important aspect of what end-users—as opposed to developers—need from these models: what is needed is *systematicity*, in a more demanding sense than the compositionality-related sense that has dominated discussions of systematicity in the philosophy of language and cognitive science over the last thirty years. To recover this more demanding notion of systematicity, I distinguish between (i) the systematicity of thinkable contents, (ii) the systematicity of thinking, and (iii) the ideal of systematic thought. I then deploy this distinction to critically evaluate Fodor's systematicity-based argument for the language of thought hypothesis before recovering the notion of the systematicity of thought as a *regulative ideal*, which has historically shaped our understanding of what it means for thought to be rational, authoritative, and scientific. To assess how much systematicity we need from AI models, I then argue, we must look to the *functions* of systematizing thought. To this end, I identify five functions served by systematization, and show how these can be used to arrive at a dynamic understanding of the need to systematize thought that can tell us *what kind of* systematicity is called for and *when*.

Keywords: artificial intelligence, thinking, language of thought, syntactic structure, semantics, compositionality, systematicity, productivity, connectionism, functions of systematization.

1. Introduction

One of the more philosophically fruitful features of the recent excitement over deep artificial neural networks is that, through their shortcomings, these models hold up the mirror to the norms that govern human thought and language. For the first time, we face systems that are deeply different from us, yet capable of producing human-like language. Our perception of their limitations is a direct reflection of the normative expectations we have towards our own thought and speech. This provides an opportunity to better understand these normative expectations and to critically reflect on them.

Here, I propose to do just that with the ideal of systematicity. The tendency of AI systems to be inscrutable black boxes has understandably led to calls for interpretability and explainability: people want the basis on which these models reach their outputs to be humanly intelligible, and they want some explanation of how the output was reached.

Yet I submit that interpretability and explainability fall short of capturing what we really want from AI—what is really at issue from the perspective of end-users as opposed to developers. The more pertinent question, for someone who is not primarily concerned to tweak the model, is to what extent AI exhibits systematicity of thought. This regulative ideal of systematicity, I argue, is fundamental to making minds and language interpretable to begin with; and it underlies our sense of what makes for *good* explanations.

This comes out clearly when considering deep neural networks trained to generate language—so-called Large Language Models (LLMs). Not only can these models yield linguistic outputs that are grammatically correct and contextually relevant; if suitably prompted, they can also generate explanations to accompany those outputs. But the limitations of LLMs reveal themselves precisely when these explanations flout the ideal of systematicity: even when an explanation is provided, the explanation may be inconsistent with other outputs of the model, or feel rationally disconnected and thus fail to cohere with them; in certain cases, it may also feel insufficiently principled, or insufficiently

parsimonious in the number of principles it invokes. These are not failures of interpretability or explainability. They are failures of systematicity.

This is occluded by the fact that the phrase "systematicity of thought" has acquired a much narrower meaning in cognitive science and the philosophy of mind and language over the past decades, especially after Jerry Fodor and Zenon Pylyshin (1988) harnessed the phrase to formulate what came to be known as the "systematicity challenge" to artificial neural networks: there is a symmetry in our cognitive capacities, they argued, in that the ability to entertain a given thought, such as "John loves Mary," implies the ability to entertain certain other thoughts, such as "Mary loves John." They referred to this as "the systematicity of thought," and contended that network architectures were ill-suited to reproduce it. If human thought is systematic, they hypothesized, this is because thought itself possesses a language-like structure. It consists in the manipulation of discrete symbols with a combinatorial syntax and compositional semantics: we recombine, in rule-governed ways, distinct mental representations that are already meaningful in isolation, and the meaning of the resulting complex thoughts can be derived entirely from the meaning of their constituent symbols together with the way these are syntactically arranged. The only way in which neural networks, which operate at the sub-symbolic level and use distributed rather than discrete representations, could ever hope to reliably reproduce this systematicity, Fodor and Pylyshin insisted, was by laboriously implementing symbolic processing within such a network architecture—which was tantamount to conceding their point.

Yet recent advances have transformed that familiar debate, making Fodor and Pylyshin's challenge lose much of its challengingness. "The remarkable progress of LLMs in recent years calls for a reexamination of old assumptions" about systematicity as a "core limitation of connectionist models," Raphaël Millière and Cameron Buckner observe (2024, §3.1). Perhaps most notably, a *Nature* article triumphantly announced that techniques such as Meta-Learning for Compositionality (MLC), which involves using metadata from previous learning episodes to improve the learning process (thereby allowing the networks to "learn

to learn"), now allow neural networks to rise to Fodor and Pylyshin's challenge (Lake and Baroni 2023). This has been hailed as a "breakthrough in the ability to train networks to be systematic" (Smolensky, quoted in Kozlov and Biever 2023, 16). With that challenge now met, or close to being met, it is time to recover the older, broader, and more demanding conception of cognitive systematicity that this "systematicity debate" has driven off the stage.

I shall proceed as follows: after reconstructing the role that a comparatively narrow notion of the systematicity of thought has played in the philosophy of language and mind as well as in the "systematicity debate" between classical and connectionist theories in cognitive science (§2), I broaden the notion by drawing a tripartite distinction between the systematicity of *thinkable contents*, the systematicity of *thinking*, and the ideal of systematic thought; I then deploy this distinction to critically evaluate Fodor's systematicity-based argument for the language-like character of thought before recovering the conception of systematic thought as a regulative ideal that has historically shaped our understanding of what it means for thought to be rational, authoritative, and scientific (§3). To assess how much systematicity we need from AI systems, I then argue, we must look to the function of systematizing thought. To this end, I identify five functions served by systematization (§4), and show how these can be used to construct a dynamic understanding of the need to systematize thought that can tell us what kind of systematicity is called for and when (§5).

2. Narrow Notions of the Systematicity of Thought

2.1 Systematicity in the Philosophy of Language and Mind

The notion of the systematicity of thought makes a number of prominent appearances in twentieth-century philosophy of language and mind, though it is fair to say that philosophers have not thought very systematically about it.

One way in which the systematicity of thought is invoked is to point out that there is a *structure* inherent in thought and language. Thus, Gareth Evans writes: "It seems to me that there must be a sense in which thoughts are *structured*. The thought that John is happy has

something in common with the thought that Harry is happy, and the thought that John is happy has something in common with the thought that John is sad" (1982, 100).

To make sense of this structure, Evans invites us to see the thought that *a* is *F* "as lying at the intersection of two series of thoughts: the thoughts that *a* is *F*, that *a* is *G*, that *a* is *H*, …, on the one hand, and the thoughts that *a* is *F*, that *b* is *F*, that *c* is *F*, …, on the other" (1982, 209). From this observation, Evans derives what he calls the *generality constraint*: "if a subject can be credited with the thought that *a* is *F*, then he must have the conceptual resources for entertaining the thought that *a* is *G*, for every property of being *G* of which he has a conception" (1982, 104).[1] This form of systematicity has been thought to be a key criterion of concept possession (Butlin 2023, §3).

Here, to say that thought is systematic is to say that there is a structure inherent in well-formed thoughts that allows systematic variants of them to be produced and understood, where systematic variation involves *permuting constituents* or, more demandingly, *substituting constituents of the same kind*.[2]

As Ludwig Wittgenstein remarks in conversation with Friedrich Waismann, it is these possibilities of permutation and substitution that give point to the internal structure of thought. Were these systematic variations not in the offing, thought might as well have no internal structure at all:

> If there were only the proposition, 'ϕa' but not 'ϕb', it would be superfluous to mention 'a'. It would suffice to write just 'ϕ'. … If 'ϕa' is supposed to be a proposition, then there must also be a proposition 'ϕb', that is, the arguments of '$\phi()$' form a system. … But does 'ϕa' presuppose 'Ψa' too? Decidedly yes. For the same consideration tells us: if there were only a single function 'ϕ' for 'a', then it would be superfluous; you could leave it out.

[1] For a defence of an unrestricted version of this constraint, see Camp (2004); for a defence of a weaker version of the constraint, see Dickie (2010); and see Travis (2015) for a critique of it that echoes Warren Goldfarb's Wittgensteinian gripes about the supposed "fixity of meaning" (1997).

[2] This sense of systematicity figures centrally in Peacocke (1992, 42), Cummins (1996, 2010), Johnson (2004), Perler (2004), and Salje (2019).

(Waismann 1979, 90)

Thus far, we therefore have the idea, first, that thought has an inherent structure permitting systematic variation; and second, that it is the *point* or *function* of that structure to permit systematic variation.

A third idea, adumbrated in the Wittgenstein-Waismann quote, is that any given thought has to be *part of a larger system* of thoughts—there could not be a thinker capable only of entertaining a single thought in isolation. To adapt an example from Andy Clark (1991): imagine being promised a robot that could think, only to discover that all the robot could do was to say "The cat is on the mat" whenever presented with a cat on a mat. Would we say that the robot could think, but only a single thought? Surely not. For as long as the robot shows no sign of being able to think other thoughts in the vicinity, such as "The mat is on the cat," it is not clear that it is really *thinking* "The cat is on the mat" at all. It is constitutive of thinking that *a* is *F* that one has the capacity to entertain other thoughts involving *a* or *F*, such as that *b* is *F*, or that *a* is not *F*, and so on. We cannot make sense of a thinker that can only entertain a single thought. An influential articulation of this point can be found in Wittgenstein's *On Certainty*: "when we first *believe* anything, what we believe is not a single proposition, it is a whole system of propositions. (Light dawns gradually over the whole.)" (1969, §141).

A fourth idea, which became particularly influential in cognitive science, is that there is what Jerry Fodor calls "a symmetry in our cognitive capacities" (1998, 26): the ability to entertain a given thought implies the ability to entertain certain other thoughts. Someone who has the capacity to think thoughts of the form *aRb* (e.g. "John loves Mary") also has the capacity to think thoughts of the form *bRa* (e.g. "Mary loves John"). However, as Fodor emphasizes, this form of "systematicity concerns symmetries of cognitive *capacities*, not of actual mental states" (1998, 26n2). Not everyone who thinks that "Humans walk dogs" also thinks that "Dogs walk humans." The point is merely that if they are able to think the former, they are also able to think the latter.

These four ideas about the systematicity of thought work well together, and one might form the impression that there is something approaching a consensus in philosophy on what the systematicity of thought amounts to. But this would be to miss the deeper divide that is masked by the superficial harmony of these four ideas. This is the divide between those who regard the systematicity of thought as an argument for semantic *atomism* and those who regard it as an argument for semantic *holism*.

On the one hand, Jerry Fodor (1998) has argued that the best explanation for the systematicity of thought is that thought must be fundamentally atomistic: made up of unstructured concepts that derive their content solely from their relation to entities in the environment rather than from their relation to other concepts. Furthermore, these atomistic concepts must be recombinable according to syntactic rules, forming complex thoughts whose content is simply a function of the concepts they are composed of together with the way they are syntactically combined. In other words, Fodor takes the systematicity of thought to reflect its *atomism*, *combinatorial syntax*, and *compositional semantics*. Together, these three features also account for the *productivity* of thought, on his account. Only a finite number of constituents need to be learned to generate an infinite variety of new thoughts—we can, in Humboldt's phrase, make "infinite use of finite means"—because these new thoughts are generated by recombining familiar constituents, producing unprecedented thoughts that can nonetheless be understood thanks to the compositionality of their semantics.

Note that on Fodor's account, the observation that the ability to entertain any one thought implies the ability to entertain others is not meant to be a holistic, conceptual point, but an *empirical* one. Fodor takes it to be "conceptually possible that there should be a mind that is able to grasp the proposition that Mary loves John but not able to grasp the proposition that John loves Mary. But, in point of empirical fact, it appears that there are no such minds" (1998, 26).

On the other hand, semantic holists take the fact that there are no such minds to reflect a conceptual point about the holistic nature of thought ascription: it is *constitutive* of genuinely grasping a thought that one is able to place it in a web of systematic relationships to other thoughts. One does not really grasp the thought that *x* is a *dog* unless one understands other thoughts that are *inferential consequences* of that thought (e.g. if *x* is a *dog*, this *implies* that it can be kept as a *pet*, and it *excludes* its being a *bird*). While Fodor's atomistic account expressly "denies that the grasp of *any* interconceptual relations is constitutive of concept possession" (Fodor 1998, 71), holistic accounts such as those of W. V. O. Quine (1951), Wilfrid Sellars (1958), or Robert Brandom (1994) maintain that grasping at least some of the inferential connections a concept stands in to other concepts *is* constitutive of concept possession. In Brandom's slogan, "one must have many concepts in order to have any" (1994, 89).[3] To genuinely count as grasping a concept, one has to grasp its role in a web of inferential connections between thoughts, for these are a crucial part of what gives determinate content to the concept in the first place.[4] Concepts would be mere labels, devoid of substantive meaning, if conceptual content were limited to the referential dimension that a concept bears to its object. Imagine being handed an *F*-detector that lights up red when and only when presented with an *F*; you could use it to sort things into *F*s and non-*F*s, but as long as you lacked any idea of what something's being *F implied*, you would not understand the *significance* of being *F* or non-*F*; it would merely be an empty label to you.[5] As Wilfrid Sellars puts the point: "It is only because the expressions in terms of which we describe objects … locate these objects in a space of implications, that they describe at all, rather than merely label" (1958, 306–7).

[3] See also Brandom (2009, 202; 2019, 113).

[4] Whether this constitutes a necessary or a sufficient part of the determination of conceptual content depends on the strength of the inferentialism one endorses; Brandom endorses *strong* inferentialism, on which inferential articulation, broadly construed (i.e. including *materially* correct inferences as well as *non*inferential circumstances and consequences of concept application), is *sufficient* to account for conceptual content (2000, 28; 2007, 657).

[5] For a similar illustration, see Brandom (2009, 202).

Since this kind of semantic holism holds that the semantic value of a complex thought cannot be computed without taking into account its relations to other thoughts, holism would seem to deny compositionality, which in turn renders it hard to see how it could account for the productivity of thought. But, as Brandom (2008, 134–36) points out against Fodor and Lepore (2002), holists can account for these features of thought just as well as atomists can. The key is to distinguish the compositionality of semantics from the recursiveness of thought. For notice that on a holistic view, the semantics of complex thoughts can be fully *determined by* the semantics of their simpler constituents without being fully *interpretable in terms of* them. That is, one cannot compute the semantic value of a complex thought without taking into account its relations to other thoughts—what thoughts it is compatible or incompatible with, and which implications of other thoughts it is compatible or incompatible with. But this non-compositional, holistic semantics at the level of complex thoughts is compatible with full recursiveness *between* levels of constructional complexity. It is this recursiveness that is really needed to ensure the productivity of thought. And this recursiveness, Brandom argues, can be preserved without subscribing to the atomistic idea that the meaning of a complex thought can be understood wholly bottom-up, independently of its relation to other thoughts.[6] If this is right, it suggests that the systematicity of thought cannot ultimately be invoked to decide between semantic atomism and semantic holism. Significantly for our purposes, however, both sides agree on its reality and importance.

2.2 The Systematicity Debate in Cognitive Science

In 1988, Jerry Fodor and Zenon Pylyshyn launched a debate in cognitive science that revolved around this notion of the systematicity of thought that came out of mid-twentieth-

[6] For a critique of this line of thought, see Fermüller (2010); for a defence of it against this critique, see Turbanti (2017, ch. 4, §4.2). I agree with Turbanti that Brandom's position only makes sense in conjunction with his expressivism.

century philosophy of language. The debate pits *classicists*, who understand cognition as rule-governed symbol manipulation running on a classic, Turing-style computational architecture, against *connectionists*, who propose an alternative computational architecture taking the form of a neural network.

While the Turing-style architecture basically consists of a central processor with some form of memory on which a string of symbols can be inscribed, neural networks have no central processor and no dedicated memory. They consist solely of a network of nodes bearing weighted connections to one another. Accordingly, they encode information not in the form of strings of discrete symbols, but sub-symbolically, in the weights or connection strengths between the nodes. By ditching the serial processing of discrete symbols for the parallel processing of sub-symbolic information distributed across the network, the connectionist architecture offers a far more brain-like model of how cognition works—and one, crucially, that frees us of the need to invoke mental symbols at all.[7] Geoffrey Hinton, a key figure in the resurgence of neural networks, described symbols as the "luminiferous aether of AI," likening them to the mythical medium that nineteenth-century physicists had postulated to account for the propagation of electromagnetic waves (Russell and Norvig 2021, 42).

In response to this wave of enthusiasm for connectionism, Fodor and Pylyshin formulated the "systematicity challenge" (Calvo and Symons 2014; Verdejo 2015). Connectionist architectures, with their sub-symbolic parallel distributed processing, could not adequately explain the systematicity of thought, Fodor and Pylyshyn maintained. We need classical symbolic models, with their rule-governed manipulation of discrete symbols, to account for the systematicity of thought—in particular, as Fodor and McLaughlin (1990) clarified, to account for it *as a matter of nomic necessity*, by explaining how the structured

[7] It should be noted that the advent of hybrid architectures such as those described in Smolensky (1990) and Eliasmith (2013) has rendered the distinction between symbolic and sub-symbolic processing less clear-cut. The two can in principle be combined to varying degrees.

nature of symbolic representations effectively *guarantees* the systematicity of thought.[8] This launched the "systematicity debate" between classicists (Fodor and McLaughlin 1990; McLaughlin 1993; Aydede 1997; McLaughlin 2009) and connectionists (Smolensky 1988, 1990; van Gelder 1990; Matthews 1994; Cummins 1996; Cummins et al. 2001).[9]

The debate was never about whether a connectionist architecture was compatible *in principle* with information processing exhibiting systematicity—it can be mathematically demonstrated that a classical symbol processor *can* in principle be implemented within a connectionist architecture. Rather, the debate was over whether a connectionist architecture is well-suited to *account for* the systematicity of thought, and whether or not it would have to implement symbolic processing in order to reliably reproduce human-like systematicity of thought.

Since Robert Hadley (1994) sought to operationalize the notion of systematicity by defining benchmarks that connectionist models would need to pass, ever more powerful neural networks have been advanced to demonstrate that they can exhibit human-like systematicity.[10] Admittedly, a difficulty has been that the apparent systematicity exhibited by neural networks may in fact be no more than the regurgitation of something buried deep within the training data—what looks like systematicity may in fact be no more than memorization. To correct for this, Ettinger et al. (2018) as well as Yu and Ettinger (2020) have developed more demanding benchmarks. Yet even neural networks trained on vast amounts of data have struggled with forms of systematicity that symbolic models handle with ease (Marcus 2001; Hupkes et al. 2020; Press et al. 2023). Recent networks such as

[8] As others have noted, however, explaining systematicity as necessitated by laws is a very strong requirement that classical architectures likely cannot meet either (Buckner and Garson 2019).

[9] See Aizawa (2003), the essays in Calvo and Symons (2014) as well as Buckner and Garson (2019) and Rescorla (2024) for thorough discussions of the debate, and see Johnson (2004) for a critique of the conception of systematicity that is supposed to underpin it. More recently, "category theory," which draws on the mathematical theory of structure, has presented itself as a third contender in the debate; see Phillips and Wilson (2016).

[10] See Hadley (1997, 2004) and Buckner (2019) for critical evaluations of some of these attempts.

GPT-4 still find systematicity to be a challenge (Lake and Baroni 2023). Nonetheless, advances involving recurrent neural networks (Frank, Haselager, and van Rooij 2009; Calvillo, Brouwer, and Crocker 2021), neurocompositional computing (Smolensky et al. 2022), and meta-learning neural networks (Lake and Baroni 2023) have enabled connectionist models to make great strides.

As a result, Fodor and Pylyshin's systematicity challenge no longer seems a real challenge to connectionism. The question now is *how human-like* the systematicity exhibited by these models has become (after all, even human beings are better at certain tasks than others in this connection—they struggle to parse sentences with complicated centre embedding, for instance).

But the key point for present purposes is that the conceptions of the systematicity of thought that have come out of mid-twentieth-century philosophy of language, and around which the "systematicity debate" revolved, remain *narrowly focused* on the constituent structure of thoughts and the compositionality of semantics. In AI research, systematicity in the narrow, Fodorian sense is sometimes simply treated as evidence that a model can handle the compositionality of meaning.[11] But with the challenge of reproducing something like systematicity in this narrow sense now largely met, it is time to recover the older, broader, and more demanding conception of cognitive systematicity that this "systematicity debate" has overshadowed.

3. Broadening the Notion of Systematicity

My aim in this section is to recover, from under the shadow of the systematicity debate, an older and richer conception of systematicity. To this end, I introduce a tripartite distinction between three senses of the phrase "the systematicity of thought." With this distinction, I propose to do two things: first, I criticize Fodor's appeal to the systematicity of thought as

[11] Lake and Baroni (2023) offer a prominent example of this tendency. But this involves a simplification that is worth resisting, as Spenader and Blutner (2007) have argued.

evidence for the language of thought hypothesis; and second, I work towards a notion of systematicity that conveys a richer sense of what kind of systematicity we need in AI.

3.1 A Tripartite Distinction

Discussions of systematicity have tended to blur an important distinction between three senses of the phrase "the systematicity of thought":

(i) The "systematicity of thought" can refer to the systematicity of *what is thought*: there is, as Wittgenstein and Evans observed, a structure inherent in the contents of thought—thought is *articulated in terms of recombinable constituents*, and the possibilities for well-formed recombinations are not random, but follow rule-governed patterns that in turn give rise to systematic interconnections between thinkable contents. This is a claim about the nature of thought itself, not an empirical observation about thinkers. Each actual thought is situated in a structured network of thinkable contents (remember Evans's "intersections"). This structure is logically independent of the capacities actual thinkers in fact possess.

(ii) However, the "systematicity of thought" can also refer to the systematicity of *thinking*. The ambiguity between the *activity* of thinking and the *object* of that activity, namely *what is thought*, is an instance of a more widespread phenomenon that has been called the "act-object ambiguity" (Alvarez 2010, 125). The systematicity of thinking refers to the patterns, interconnections, and regularities discernible in the activity of thinking as performed by actual thinkers: which thoughts they can entertain, whether the capacity to entertain some thoughts entails or tends to entrain the capacity to think others, and whether the capacity to draw certain inferences entails or tends to entrain the capacity to draw other inferences.

(iii) And finally, the "systematicity of thought" can refer to a *regulative ideal* governing what thought should be like: the ideal whereby thought *should* exhibit systematicity. Instead of describing some property inherent in thinkable contents, or some

property that our thinking *de facto* already tends to display, the phrase then articulates an aspiration that needs to be realized by *systematizing* thought, that is, rendering it *more* systematic than it already is. Like interpretability and explainability, the systematicity of thought in this sense is a desirable property of thought that it may display to varying degrees.

This gives us a tripartite distinction between (i) the systematicity of *thinkable contents*, (ii) the systematicity of *thinking*, and (iii) the ideal of systematic thought.

3.2 Reevaluating Fodor's Argument from the Systematicity of Thought

Recognizing this tripartite distinction allows us to better understand and reevaluate Fodor's argument from the systematicity of thought to the language of thought hypothesis. In light of this tripartite distinction, it emerges that Fodor's argument from the "systematicity of thought" is really an argument from the systematicity of *thinking*. He observes that, as a matter of empirical fact, thinkers capable of entertaining thoughts of the form *aRb* also have the capacity to entertain thoughts of the form *bRa*. But this "symmetry," as he calls it, is a pattern in the capacities of actual thinkers. Whenever thinkers display the former capacity, they also display the latter. Likewise, Fodor notes that thinkers capable of inferring "*p*" from "*p* and *q*" are also capable of inferring "*m*" from "*m* and *n*." But again, this is, in the first instance, a pattern in our thinking. Our inferential capacities, like our capacities to entertain certain thoughts, come in bundles, so that whoever possesses one capacity also possesses the other. Yet this is in the first instance an observation about the constant conjunction of certain skills, not one about the inherent structure of what is thought.

In a second step, Fodor then argues that the observable systematicity of thinking requires explanation. Why is it that certain cognitive capacities go together in this way? His answer is that the systematicity of thinking must reflect a systematicity in thought itself: the thinkable contents themselves must be articulated in terms of recombinable constituents. So far, so uncontroversial.

But Fodor then makes a crucial further step: he takes the systematicity of thinking as evidence for the reality of a private mental language: *Mentalese*, the language of thought, which he imagines to consist of symbols or discrete mental representations with syntactic and semantic properties closely resembling those of natural languages.

Fodor's argument is therefore not really an argument *from* the systematicity of thought; it is better described as an argument from the systematicity of thinking *to* the systematicity of thought—where the latter is taken to refer to the combinatorial syntax and compositional semantics of a private language of thought.[12]

Notice, however, that this argument draws part of its plausibility from the fact that it trades on a further equivocation: it is *prima facie* compelling to say that the systematicity of *thinking* is based on the underlying systematicity of *thought*, and that thought must therefore have some inherent structure; but the familiar phenomenon we call "thought," whose public manifestations we encounter in everyday linguistic and non-linguistic behaviour, and which is widely agreed to be analysable into discriminable—though not necessarily detachable (Ryle 2009, 192)—concepts, is one thing; the inner, private language of thought that Fodor invites us to postulate is quite another. Even after we have disambiguated thinking from thought, therefore, it seems that the systematicity of thought still *comes in twice* in Fodor's story: an inner, private representational system is postulated to explain the systematicity of thinking, which in turn explains the systematicity of publicly manifestable thought.

Yet, from the fact that thought—the publicly manifestable product of our thinking capacities—is articulated in terms of recombinable constituents, it simply does not follow that we must each think in terms of a private system of mental symbols.[13] Indeed, Evans

[12] This comes out clearly in Fodor and Pylyshin (1988, 26n25).

[13] A further difficulty with this line of argument is that even if the systematicity of thought were necessary for the systematicity of thinking, it still would not be sufficient to explain it. Additional facts about cognitive architecture, learning history, and environment are needed to explain why thinking follows the particular patterns we observe—or why they sometimes perplexingly fail to be realized, as when Google's Gemini 1.5 can tell one that the mother of Tom Cruise is Mary Lee Pfeiffer, but then proves unable to tell one who the son of Mary Lee Pfeiffer is.

himself is careful to block this inference when he remarks, right after asserting that thoughts are structured: "This might seem to lead immediately to the idea of a language of thought … However, I certainly do not wish to be committed to the idea that having thoughts involves the subject's using, manipulating, or apprehending symbols" (1982, 100–101). There is a danger here of overintellectualizing human thought, and of modelling the structure of subpersonal processing too closely on the structure of natural language.

Instead, we should explain the sense in which thoughts are structured "in terms of their being a complex of the exercise of several distinct conceptual *abilities*," Evans suggests; "someone who thinks that John is happy and that Harry is happy exercises on two occasions the conceptual ability which we call 'possessing the concept of happiness'" (1982, 101). This in effect reverses the direction of explanation. Thought is systematic because our conceptual abilities are systematic.[14]

All the attention devoted to the systematicity of thought in senses (i) and (ii) has, however, overshadowed an older, broader, and more demanding sense of the systematicity of thought, namely (iii): the systematicity of thought as a regulative ideal. As recent advancements in the training of artificial neural networks have arguably put to rest the debate around whether connectionist architectures could meet Fodor and Pylyshin's challenge, this conception is now worth pulling to the fore again.

3.3 Recovering the Systematicity of Thought as a Regulative Ideal

Moving further back in the history of reflection on the systematicity of thought, it quickly becomes clear that systematicity has been understood, at least since the ancient Greeks, not as a property that human thought already has by default, but as a central regulative ideal

[14] This need not amount to an *identification* of concepts with abilities. For the reasons enumerated in Glock (2006, 2009a, b, 2010a, b, 2020), this would be too simple; concepts occur or are involved in thoughts in a way in which abilities are decidedly not. Nonetheless, to possess a concept is to possess certain discriminatory, classificatory, and inferential abilities. That much is granted even by Fodor: "*having* a concept is: *being able* to mentally represent (hence to think about) whatever it's the concept of" (2003, 19, emphasis added).

articulating what human thought *should strive for*. In calling it a *regulative* ideal, I mean to underscore that it articulates an ideal that is typically not fully realized, but that nonetheless pervasively informs and guides our thinking—much as theorizing in the natural sciences is guided by the regulative assumption of the systematic unity of nature, even if our current theories do not yet offer a systematically unified theory of everything.

Systematicity in this sense requires more than being *structured* in the minimal sense of being articulated in terms of recombinable constituents. It requires being structured *in a particular way*—namely so as to embody order and harmony. There are multiple dimensions to this, and so the ideal of systematic thought bundles together several demands:

- the demand for *explicitness*: the discursive articulation of what is implicit;
- the demand for *consistency*: the absence of contradictions;
- the demand for *coherence*: interconnectedness through relations of rational support;
- the demand for *comprehensiveness*: avoidance of lacunae;
- the demand for *principledness*: subordination of the particular to the general, such as laws or principles of which the particular is an application;
- the demand for *parsimony*: economy of laws or principles.

Explicitness, consistency, coherence, comprehensiveness, principledness, and parsimony—these are *dimensions* of systematicity, which thought can exhibit to a greater or lesser degree. Accordingly, the ideal of systematic thought relies on a gradable conception of systematicity. Further dimensions can and have been added to the list, but these tend to be either domain-specific or else subsumable under these "big six."[15]

It will be clear already from this list, however, that these dimensions can be *antagonistic*, i.e. not all fully co-realizable without trade-offs—comprehensiveness may only be achievable at the expense of parsimony, for example.[16] Just as designing a car involves striking a

[15] For more inclusive but, to my mind, less generalizable lists, see Rescher (1979, 10–11; 2005, 25–26) and Hoyningen-Huene (2013, 35–36).

[16] See Brun (2020, 950).

reasonable balance between speed and safety, systematization involves striking a reasonable balance between various dimensions of systematicity.[17]

This ideal of systematic thought has influenced Western thought at least since the ancient Greeks. Archimedes' systematization of statics and Euclid's systematization of geometry provided a lasting paradigm of systematic thought, emulated by thinkers as different as Hobbes, Spinoza, and the early Wittgenstein.

But my point is not simply that the ideal of systematicity has a long and distinguished history. My point is that this history can indicate important *functions* of systematization. Our sense of how far opaque AI models offend against the ideal of systematic thought, and of what kind of systematicity we really want from those models, is only as clear as our understanding of the original rationales for systematization in other connections.

In the next section, I therefore propose to highlight five functions of systematization: its role in the constitution of objective experience; its role in enabling understanding; its role as a criterion of acceptability; its role in instigating critical revision; and its role in exposition, persuasion, and retention.

4. Five Functions of Systematization

The first and most fundamental function of the ideal of systematicity is to drive the integration of different modalities to form the experience of an objective world. We can call this the *cognitive integration* function. This is what led Enlightenment thinkers to cast the imperative to systematization as a demand *inherent in reason or rationality itself*. Systematicity, one historian of German idealism notes, was the "unquestioned answer to the question of the nature of reason's fundamental demand" (Franks 2005, 3).[18] Striving for

[17] See Rescher (1979, 17).

[18] On Kant's aspiration to cognitive systematization, see Rescher (2000, 64–98), Kitcher (1986), Guyer (2003, 2005), Abela (2006), and Ypi (2021). The theory of systematization really came into its own in the

systematicity was thought to be *constitutive* of rationally apprehending the world: without systematization, no objective experience. Were one completely indifferent to the consistency and coherence of one's perceptions, judgements, and intentions, it is doubtful that these would still have any determinate content or representational purport at all. After all, someone who perceived a stick as *rigid and straight* when out of the water and as *rigid and bent* when half-submerged, but remained utterly indifferent to the inconsistency between these two appearances, would not be treating them as two modes of appearance *of one and the same object* at all.[19] The ideal of systematicity informs the very process whereby the sensory manifold becomes the experience of an objective world. Thus, Kant observed: "In accordance with reason's legislative prescriptions, our diverse modes of knowledge must not be permitted to be a mere rhapsody, but must form a system" (1929, A832/B860).

Kant was not writing about multimodal neural networks, but the point he was making is general enough to apply to them as well. The systematic integration of our information processing across various modalities requires more than a constituent structure; it requires striving for consistency and coherence across these modalities. This is something that becomes particularly evident in Vision-Language-Action Models (VLAMs). Unlike simple Large Language Models, which are confined to manipulating text, VLAMs are embodied Large Language Models that are capable of receiving visual input from the world and effecting changes in it through their own movements. This requires them to fit together images, words, and data running to and from their internal sensors about the angle and position of their actuators (such as grippers or hands). That requires systematic cognition in

---

Enlightenment, with Wolff's *De differentia intellectus systematici & non systematici* (1729) and Lambert's posthumously published fragments on *Systematologie*. Condillac's *Traité des systèmes* (1749) then distinguished "bad" systems resting on mere hypotheses and speculation (e.g. Pre-Socratic systematizations of nature) from "good" systems resting on experience (e.g. Newton's systematization of celestial mechanics). Similarly, D'Alembert's *Discours préliminaire* (1751) rejected the *esprit de système* in favour of the *esprit systématique*: Procrustean efforts to fit the world into preestablished metaphysical systems were to make way for careful studies of phenomena that distilled systematic principles from them.

[19] For a detailed discussion of this example, see Brandom (2019, 75–80).

a more demanding sense than was at issue in the systematicity debate. It requires a multimodal integration of visual data processing with natural language processing and action execution. A VLAM cannot afford for its different modes of information processing to be a mere rhapsody; they must form a system.

One of the things that is often perceived to be lacking in LLMs is a "grounding in the objective world" or a certain "common sense" about how things actually behave in the world. That can be interpreted as being fundamentally a complaint about a lack of systematic integration between different modalities. And even just within the sphere of natural language processing, the next-word prediction approach that underlies LLMs makes them, at least at base, indifferent to more demanding norms of systematicity such as consistency and coherence across different conversational threads. As Cameron Buckner notes:

> Some of the most disconcerting weaknesses of these models' performance pertain not to their ability to solve particular grammatical, logical, or mathematical problems, but rather their tendency to meander incoherently in longer conversations and their inability to manifest a coherent individual perspective. (Buckner 2023, 283)

An LLM may sound like an erudite human in one thread, but the illusion breaks down once a different set of prompts in a new thread produces a new web of sentences that lacks coherence with, or even flagrantly contradicts, what it asserted in an earlier thread. This experience of a *disjointed mind* erodes the impression of mindedness *tout court* (in what sense does one still believe that *p* if one also believes that *not-p*?).

Even within one thread, the systematic integration of assertions is something that next-word prediction alone does not inherently aim towards. This is sometimes put in terms of the complaint that LLMs cannot plan ahead and ensure that the next word fits not just with preceding words, but with the sentences it is going to write thereafter. In this sense, an LLM does not know what it thinks until it has read what it is going to write.

Developers have found an ingenious way of addressing this lack of systematicity by approaching the problem in multiple steps: they have the output of one "AI agent" checked

for consistency and certain forms of coherence by another "AI agent," though both agents may be running on the same LLM (Lewis and Sarkadi 2024; Shinn et al. 2024). This can be thought of as an operationalization of *reflection* in the most basic sense. But ongoing attempts to do this by chaining together multiple AI agents underscore the point that what we want from AI systems is not limited to interpretability and explainability. We expect minds to be *unified*—and that means *systematic* in the demanding sense I have been emphasizing.

This is connected to the second function of systematization, which is to render thoughts intelligible in terms of their interconnections to other thoughts. We can call this the *hermeneutic* function of systematization. Systematicity is fundamental to sense-making and understanding: we do not really understand a belief, a claim, or even an action as long as we cannot see how it relates to the judgements from which it itself follows, and to the judgements that are implied or ruled out by it. These inferential connections to other judgements are part of what gives a judgement a determinate content in the first place.[20] Integrating a judgement into a network of systematic relationships is thus part of what it means to understand it. It is only interpretable as determinately contentful insofar as we can situate it within what Wilfrid Sellars, drawing on Kant, called "the space of reasons" (1997, §36).

If this is right, it means that where the interpretability of an AI system's output by end-users is concerned, the demand for a certain degree of systematic integration is not a *separate* demand over and beyond the demand for interpretability, but really encompasses the demand for interpretability *through* systematic integration. It is in this sense that, as I put it in the introduction, the regulative ideal of systematicity is fundamental to making minds and language interpretable to begin with.

Third, integrability within a system of thoughts acts as a *criterion* for the acceptability of judgements. This is the *epistemological* function of systematization. That point was

[20] See Rescher (2005, 11–13) and Brandom (1994, 2000, 2009).

emphasized by British Neo-Hegelians such as F. H. Bradley, who sought to establish "the claim of system as an arbiter of fact" (1909, 489). But intimations of this idea can already be gleaned from Plato's insistence, in the *Theaetetus* (201c–210d), that a claim needs a *logos* (a reason) to count as knowledge; or from Aristotle's thesis, in the *Metaphysics* (I, 982a) and throughout the *Posterior Analytics*, that *episteme* is knowledge of certain principles and causes (a thesis that Thomas Aquinas influentially rendered as *sapientis est ordinare*—"the task of the wise is to systematize").[21] More recently, this is also what Nicholas Rescher has extolled as the main purpose of cognitive systematization: it enables "quality control of knowledge claims" (2005, 27). Integrability within a system offers a *criterion* by which to separate genuine *knowledge* from mere opinion.

This has been thought to be one of the principal reasons why academic disciplines strive to be even *more* systematic than ordinary thought. The degree of systematicity of a body of thought is widely taken to reflect the degree of its seriousness and authority. In other words, a high degree of systematicity has become the hallmark of *science* in the broad sense covering everything taught at a research university. This is, indeed, the central finding of Paul Hoyningen-Huene's *Systematicity: The Nature of Science* (2013).[22]

The epistemological function of systematic integrability as a criterion of acceptability is crucial to our dealings with AI models, because the basic problem of trust that we face when confronted with their output is, at heart, a problem about whether should *accept* their output. And if a key criterion of acceptability is whether a claim can be systematically integrated into the wider system of things we take to be true, it is only natural that we should want the outputs of these models to be *systematic* in the sense of being explicitly rationally supported by further explanations or justifications—not just any explanations or

[21] See Aquinas (1969, lec. 1, 1).

[22] Hoyningen-Huene (2013, 35–36) identifies nine dimensions along which scientific knowledge tends to be more systematic than ordinary knowledge, and illustrates them using examples from the history of the natural sciences and mathematics. For an overview of the largely sympathetic reception of the book's thesis in the philosophy of science, see Bschir, Lohse, and Chang (2019).

justifications, however; a *good* explanation or justification is one that is consistent with other explanations and justifications and coheres with them through relations of rational support; moreover, these explanations and justifications should not seem *ad hoc*, or multiply to become as numerous as the items they are invoked to support, but should score high along further dimensions of systematicity, such as comprehensiveness, principledness, and parsimony.

By embedding the demand for explainable AI within the broader demand for systematicity, we thus capture the important point that getting an AI model to reliably accompany its output with explanations or justifications is not enough. Even highly principled explanations or justifications are not enough. We also want our explanations and justifications to be *unified* as far as possible. This means not only that the principles should, as far as possible, be consistent and coherent, but that we value *economy* of principles. We do not want an endlessly inventive AI that assures us: "If you don't like my principles, I have others."

Fourth, systematization also performs an important *critical* function. Making the effort to systematically integrate various judgements can be a way of bringing to light how the systematic presuppositions or implications of some judgements *conflict with* the systematic presuppositions or implications of other judgements. The imperative to systematize then instigates a critical revision that might not have happened otherwise. It is often only through the effort to validate one's judgements and decisions by tracing them to more general, consistent, and coherent principles that one becomes aware of latent tensions in one's outlook, and can lean on one's principles to overturn one's prejudices.[23]

In interactions with AI models we do not fully trust, this critical function of

[23] Systematicity in this revisionary role figures prominently in the work of certain ethical theorists such as Ross (1930), Hare (1952, 1972, 1989, 2002), and Kagan (1989), which has provoked a debate over the extent to which, in ethics, rationality should be understood in terms of systematicity at all; see Berlin (2002, 2013b, a), MacIntyre (1978, 1988, 2013), Taylor (1985, 1989), Williams (1981, 1985), Stocker (1990), Dancy (1995, 2004), Wolf (1982, 1992, 2007, 2010), Chappell (2015), and Geuss (2020).

systematization can be performed in two directions: an AI model capable of systematization can help one think through the implications of one's own views and awaken one to inconsistencies in them; or the critical leverage generated by systematization can be applied to the AI model itself. By pushing an AI model to explicitly integrate its output into a wider network of thoughts and demonstrating the systematicity of that network, one renders inconsistencies and other flaws glaringly obvious. Systematization then enables others to verify that some judgement or decision is indeed based on the right kind of supporting considerations. As Cueni and Queloz (2021) have argued, this is why people in positions of public authority—such as judges, government commissions, or hospital ethics committees—are expected not just to hand down decisions, but to make discursively explicit how these decisions follow from more general principles that have been consistently and coherently applied. That is part of what it means for the decision-making to be fair and to treat like cases alike. Systematicity provides accountability, and enables those at the receiving end of these decisions to *ascertain* their fairness.

Rendering fairness verifiable through systematicity is clearly a desirable feature in AI models. As the fast-growing literature on "fair-AI" brings out, part of the anxiety surrounding the opacity of these models stems from the worry that they might be unfair by not treating like cases alike.[24] This worry is usually articulated in terms of a demand for explainable AI. But the present argument suggests that this demand can be subsumed under a more general demand for *systematic* AI. Systematicity facilitates just the kind of critical oversight that clamours for explainability tend to be after.

Moreover, an AI capable of such systematization would *ipso facto* be able of *self*-critique. The critical function of systematization would thus not only allow us to use AI to reflect critically on our own outlook, but also allow AI models to critically improve their output, including with regard to whether they were treating like cases alike.

Fifth, systematization performs an underappreciated *pedagogical* function, facilitating

[24] For a recent "state-of-the-art" of fair-AI methods and resources, see Alvarez et al. (2024).

exposition, persuasion, and retention. By having something laid out in terms of a perspicuous system, we gain a sense of its structure—not in the narrowly focused, syntactic sense, but in the broader sense of how a complex array of thoughts can be *organized into a systematic order*. This renders a complex array of thoughts much easier to convey and internalize; and it also has a protreptic effect, rendering the array of thoughts more persuasive by displaying its inner consistency and coherence. This protreptic effect relies on the fact that we tend to regard systematicity as a criterion of truth and a hallmark of authoritativeness. And, last but not least, systematization renders a complex array of thought much easier to memorize. Systematicity is a powerful mnemonic device.

This pedagogical function of systematization is evident in the history of the notion of cognitive systematicity. When the term "system" begins to be applied to systems *of thought* during the Renaissance, this is done with a view to facilitating exposition, persuasion, and retention. Of course, the *term* "system" goes back to the ancient Greek terms *systema* and *systematikos*, but these were not initially used to express the *concept* at issue here. *Systema* comes from *syn-histemi*, "to (make to) stand together," and the term was originally applied not to a constellation of judgments exhibiting systematicity, but to flocks of animals, formations of soldiers, or composite political units.[25] And while the Stoics began to use the term in a technical sense to refer to the *systema mundi*,[26] the orderly whole of heaven and earth, they still used it to refer to the systematicity of the world rather than of thought.[27] It was only in the sixteenth century that Protestant theologians, reacting to the confusion about what followers of the new faith were supposed to believe, began to publish synoptic expositions of the new theological doctrines, and referred to them as "systems" (Ritschl 1906, 16). Philosophers followed suit with works such as Keckerman's *Systema logicae* (1600) and

[25] Though Plato also speaks of the Pythagorean idea of a "system" of intervals between notes in the *Philebus* (17d).

[26] See the fragment from Chrysippus in Arnim, *Stoicorum Veterum Fragmenta* (1964, vol. 2, p. 168, l. 15).

[27] This contrast between *objective* and *cognitive* notions of systematicity is emphasized by Rudner (1966, 89), whereas Marchal (1975) highlights their common origin.

Timpler's *Metaphysicae Systema Methodicum* (1604). But the rationale for these systematizations of theology, logic, and metaphysics was primarily a pedagogical one.

To summarize, systematizing thought can perform five functions:

1. The *cognitive integration* function, whereby the integration of different modalities helps one form the experience of an objective world;
2. The *hermeneutic function*, whereby thoughts are rendered intelligible in terms of their interconnections to other thoughts;
3. The *epistemological function*, whereby integrability within a system of thought acts as a *criterion* for the acceptability of judgements;
4. The *critical function*, whereby the systematization of thought instigates critical revision;
5. The *pedagogical function*, whereby the systematization of thought facilitates exposition, persuasion, and retention.

In light of these five functions, we can see that it is worth wanting more than interpretability and explainability from AI models. Developers may primarily need to be able to interpret the inner workings of these models and to be able to explain on what basis they reached their output; but end-users of these models have different needs. They need these models to exhibit a certain degree of systematicity, not just in the thin sense of correctly processing recombinable constituents, but in the thicker, more demanding sense of systematically integrating their thoughts to form an explicitly and parsimoniously justifiable, consistent, coherent, and comprehensive whole.

Much of what is felt to be lacking in the present instantiations of these models is thus not best captured simply in terms of interpretability and explainability. From the end-user perspective, it is really systematicity that is called for, and interpretability and explainability can largely be subsumed under this broader demand. However, as the next and final section will argue, the five functions of systematization can also provide a guiding sense of *when* and *how* AI models need to systematically integrate their output.

5. A Dynamic Understanding of the Need for Systematicity

Once we recognize that the regulative ideal of systematicity answers to a practical need for systematization to discharge certain functions, the question "Should AI be systematic?" takes on a different shape. It no longer looks like a binary yes-or-no question, because we can have *more* or *less* of a need for systematization; and the question no longer appears answerable in the absolute, because the extent to which we have a need for systematization depends on the concrete practical context from which the need arises. The need for systematic AI then appears *scalable* and *context-sensitive*—a need that grows out of, and varies with, AI models' context of application.

Accordingly, we can derive a *dynamic* understanding of the need for systematic AI by conceptualizing this need as a function of the following three parameters:

> [Which AI model] needs to discharge [which function of systematization] for [which human agents]?

The first parameter registers the fact that whether AI needs to be capable of systematization depends notably on which kind of AI model we are talking about—a network trained to detect breast cancer plays a very different role in human affairs from one trained to predict recidivism, or to play Go, or to teach trigonometry lessons.

The second parameter registers the fact that how much systematization is needed in a given context, and along which dimensions (i.e. discursive explicitness, consistency, coherence, comprehensiveness, principledness, or parsimony), depends on which functions of systematization need to be discharged in that context. Some functions require a greater degree of systematization than others; and while some call above all for consistency, others call additionally for principledness; some may give more weight to comprehensiveness, while others may above all require parsimony. What weight we give to these different

dimensions of systematization should depend on the function that the systematization is meant to serve.

The third parameter, finally, registers the fact that the need for systematic AI also depends on what kind of human agent is dealing with its output. Is it the AI model's own developers, or is it end-users? Are these end-users school children or math teachers? Defendants or judges? What human agents need from an AI model will vary significantly with the role they occupy.

Depending on which values these three parameters take, the resulting understanding of the need for systematic AI will be more or less demanding. If we ramp up all three parameters to maximum scope, for example, this gives us the following, highly demanding conception of systematicity in AI:

> Every AI model needs to discharge all five functions of systematization for every conceivable human agent.

The need for AI to display systematicity of thought would then be ubiquitous. It would also take a highly demanding form, because the systematicity would have to be such as to fulfil all five functions—and not just for a specific kind of end-user, or for a specific group of people at a certain point in history, but for *every conceivable* human agent. Agents occupying the same end-user role, or a certain sociohistorical perspective, share, in virtue of that fact, certain concepts, capacities, interests, values, intuitions, and background assumptions. But the wider the range of the systematization's addressees, the less shared material there is for the systematization to rely on. What do all conceivable human agents have in common? The notion of a systematization of thought that would make sense to any conceivable human agent is so thin as to be almost completely indeterminate.

At the other extreme, setting all three parameters to their minimum scope effectively obviates the need for systematic AI:

> No AI model needs to discharge any of the functions of systematization for any human agent.

To understand the need for systematic AI as scalable and context-sensitive is to understand it, first, as lying *on a scale somewhere between* these two extremes; and second, as lying *at a different place* on that scale *depending on context*—in particular, on the value of the three parameters.

My suggestion is that we should treat the functions of systematization as the apex in this triangle of parameters. That is, our sense of *when* and *how* AI models need to systematize should be guided by our sense of the *functions* that need to be discharged in a given context. Systematicity should be proportional to the need for systematization, and the need for systematization reflects the need for certain functions to be discharged.

Thus, if we ask which AI models need to discharge the cognitive integration function, it is not clear that every AI model is subject to that need to the same degree. That need is particularly acute for multimodal models, and VLAMs especially, which have to fit together images, words, and data running to and from their internal sensors. But a vision model that is not embodied can afford to be much less systematic, as can a language model on its own. If there is a need for cognitive integration here, it comes from *us*—it is *our* need for cognitive integration, rather than that of an artificial intelligence, that calls for a certain degree of systematization even then. In other words, it is the third parameter—*for whom* the models need to systematize—that generates a demand for systematization even within models that are not multimodal. At the same time, the *kind* of systematization required to discharge this function will above all foreground consistency; some degree of coherence may also be helpful, but explicitness, comprehensiveness, principledness, and parsimony do not seem called for by the need for cognitive integration alone. This function is one that calls for a fairly minimal degree of systematization.

Insofar as the hermeneutic function needs to be discharged, by contrast, coherence becomes far more important, since mere consistency is often not yet enough to *understand* something. Whether the kind of coherence is explanatory or justificatory, i.e. whether it renders intelligible what causally led to an AI model's output or rather offers a *post hoc* rationalization of the output, will depend on the kind of addressee involved, and whether they are more interested in understanding how the output was reached or in whether they should act on it. In either case, the model will have to be capable of explicitly articulating propositions that its output coheres with (these could be mathematical propositions). Some measure of principledness may also be called for, since we often need to understand the particular's relation to the general in order to properly understand it. But comprehensiveness and parsimony will not necessarily play a large role where the hermeneutic function is concerned. It is only certain forms of understanding—of the kind sought in fundamental physics, for example—that insist on a parsimonious basis that is also comprehensive in its explanatory scope.

Accordingly, there is more of a pressure on AI models used in fundamental research to trace their output all the way back to first principles—since highly general principles and their comprehensive applicability to the particular are precisely what people engaged in fundamental research are trying to uncover. But this hermeneutic function is by no means one that every AI model has to discharge. AI models used to summarize or simplify documents, for example, will be evaluated only according to whether their output faithfully distilled what they were given as input. No additional commentary or systematic integration will be needed, as the model is being used not to *add* to what we understand, but to accurately compress it for certain purposes—and the criterion of accuracy is externally given by the original input.

When the function to be discharged is epistemological, by contrast, the criterion of accuracy will, by definition, not be externally given: systematic integration is then needed precisely to help us tell whether the output is one we should accept. Here also coherence is

needed in addition to consistency, because consistency alone is typically not yet enough to decide whether to accept something. A model's ability to make discursively explicit what rationally supports its output will then also be conducive to discharging that function, as will comprehensiveness, principledness, and parsimony to a greater or lesser degree, depending on whether the users bring ordinary or scientific standards of acceptability to the situation (striving for a particularly high degree of systematicity is, as we saw, a hallmark of science). Yet if we ask *which* AI models must discharge this epistemological function and *for whom*, it will only be those where there is even a question of us accepting the output as true or justified. And this is only the case when the human users are in the business of fact-finding at all—which is by no means the only purpose to which people put these models. An AI that dutifully demonstrated how to derive everything it said from first principles would be useless for most creative or conversational purposes. (As philosophers soon discover at cocktail parties, there is such a thing as systematizing to the point of being a bore.)

When it is the critical function that needs to be discharged, however, principledness comes to the fore, since tracing a judgement to more general principles, and searching for inconsistencies either between these principles or in their application, is one of the main ways in which we acquire critical leverage over judgements. As the literature on reflective equilibrium shows, this works in both directions, in that particular judgements might lead one to revise principles just as principles might lead one to revise particular judgements.[28] Moreover, the consistency and coherence of the principles themselves becomes an important consideration in this employment of systematization for critical scrutiny, since the consistent and coherent application of consistent and coherent principles is the mark of non-arbitrary power—and one of the animating anxieties prompting critical scrutiny of AI models is precisely the worry that their output might be arbitrary.

[28] See Elgin (1983, 1996, 2017) for epistemological applications of the reflective equilibrium model. For a recent reevaluation, see Beisbart and Brun (2024).

The same logic applies to cases in which AI models are being used to critically reflect on one's own beliefs and principles. Here too, a comparatively high degree of systematization and principled reasoning will be required to uncover tensions that were buried deep enough not to be obvious from the start. Nonetheless, it would require additional and contested philosophical assumptions for this to lead to the conclusion that AI models should be able to systematize all the way to an axiomatized theory. Absent such assumptions, a reflective equilibrium conception on which they achieve critical leverage by systematizing some of the way, but not all of the way to a small handful of axioms, offers a less contested model of what kind of systematization we are after for critical purposes.

Finally, the pedagogical function of systematization reminds us that many AI models will be used not to discover new things or question old things, but simply to convey what is already widely accepted. This applies paradigmatically to AI models optimized for teaching, where systematization plays a crucial role in organizing vast amounts of information to render it learnable, compelling, and memorable. But it applies also to AI models acting as search engines, since these gain an edge over traditional search engines precisely by laying out information in more perspicuous, systematic ways, and by making it easy to ask follow-up questions to prompt a deeper systematic integration of the initial output.

Thus, by understanding the regulative ideal of the systematicity of thought as answering to practical needs, we can conceive of it as scalable and context-sensitive; and by taking the functions of systematization as a guide to *which* models should systematize, *how* they should systematize, and *for whom*, we get a dynamic understanding of the need for systematicity that yields, not a rigid, one-size-fits-all benchmark for all AI models to meet, but rather a more nuanced ideal that can be discriminatingly pursued and tailored to human needs in particular contexts.

6. Conclusion

The systematicity of thought has long been a prominent topic in philosophy, cognitive science, and AI research. But as we have seen, prevailing definitions do not do justice to the complexity of the phenomenon, tending to characterise it only as far as this serves a particular agenda. As a result, the systematicity of thought is consistently understood in a highly narrow way, which no longer merits the exclusive attention it has received now that artificial neural networks are coming increasingly close to achieving what Fodor and Pylyshin predicted they never would.

I have sought to recover, from the deeper history of philosophy, a broader, richer, and more demanding conception of systematicity that has to some extent been obfuscated by the recent systematicity debate. I have situated this broader conception in relation to that debate using the distinction between the systematicity of thinkable contents, the systematicity of thinking, and the ideal of systematic thought.

After turning that tripartite distinction against Fodor's systematicity argument for the language of thought hypothesis, I used it to recover the broader notion of the systematicity of thought as a regulative ideal. I then proposed five functions that the pursuit of this ideal performs, and suggested that by keeping an eye cocked on these functions, we could derive a guiding sense of when, how, and for whom AI models should systematize.

This programmatic sketch still leaves a great deal of more fine-grained work to be done. But its overarching aim has been to show that it is worth looking beyond the buzzwords "interpretability" and "explainability." While important, these remain too narrow and overly focused on features that are more pertinent to developers rather than end-users of AI models. In the time-honoured notion of systematicity, we already have a powerful regulative ideal that effectively *underlies* our sense of what is interpretable and what is a good explanation, because that ideal has shaped our conceptions of what it means for thought to be rational, authoritative, and scientific. Aiming for systematic AI therefore promises to give us a far richer sense of what future AI models ought to be capable of. In the meantime,

recognizing that the ideal of systematicity informs our perception of these models can elucidate the respects in which they still strike us as inadequate. For, unlike interpretability and explainability, the ideal of systematic thought gets to the heart of what it means for artificial intelligence to come across as intelligent at all.

Bibliography

Abela, Paul. 2006. 'The Demands of Systematicity: Rational Judgment and the Structure of Nature'. In *A Companion to Kant*. Edited by Graham Bird, 408–22. Oxford: Blackwell.

Aizawa, Kenneth. 2003. *The Systematicity Arguments*. Dordrecht: Springer.

Alvarez, Jose M., Alejandra Bringas Colmenarejo, Alaa Elobaid, Simone Fabbrizzi, Miriam Fahimi, Antonio Ferrara, Siamak Ghodsi, Carlos Mougan, Ioanna Papageorgiou, Paula Reyero, Mayra Russo, Kristen M. Scott, Laura State, Xuan Zhao, and Salvatore Ruggieri. 2024. 'Policy Advice and Best Practices on Bias and Fairness in AI'. *Ethics and Information Technology* 26 (2): 31.

Alvarez, Maria. 2010. *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford: Oxford University Press.

Aquinas, Thomas. 1969. *Sancti Thomae de Aquino Opera Omnia: Sententia Libri Ethicorum, Vol. 1*. Rome: Ad Sanctae Sabinae.

Arnim, Hans Friedrich August von. 1964. *Stoicorum Veterum Fragmenta*. Stuttgart: B. G. Teubner.

Aydede, Murat. 1997. 'Language of Thought: The Connectionist Contribution'. *Minds and Machines* 7 (1): 57–101.

Beisbart, Claus, and Georg Brun. 2024. 'Is There a Defensible Conception of Reflective Equilibrium?'. *Synthese* 203 (79): 1–27.

Berlin, Isaiah. 2002. 'Two Concepts of Liberty'. In *Liberty*. Edited by Henry Hardy, 166–217. Oxford: Oxford University Press.

Berlin, Isaiah. 2013a. 'The Decline of Utopian Ideas in the West'. In *The Crooked Timber of Humanity: Chapters in the History of Ideas*. Edited by Henry Hardy, 21–50. Princeton: Princeton University Press.

Berlin, Isaiah. 2013b. 'The Pursuit of the Ideal'. In *The Crooked Timber of Humanity: Chapters in the History of Ideas*. Edited by Henry Hardy, 1–20. Princeton: Princeton University Press.

Bradley, F. H. 1909. 'Coherence and Contradiction'. *Mind* 18 (72): 489–508.

Brandom, Robert. 1994. *Making It Explicit. Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.

Brandom, Robert. 2000. *Articulating Reasons*. Cambridge, MA: Harvard University Press.

Brandom, Robert. 2007. 'Inferentialism and Some of its Challenges'. *Philosophy and Phenomenological Research* 74 (3): 651–676.

Brandom, Robert. 2008. *Between Saying and Doing*. Oxford: Oxford University Press.

Brandom, Robert. 2009. *Reason in Philosophy: Animating Ideas*. Cambridge, MA: Belknap Press of Harvard University Press.

Brandom, Robert. 2019. *A Spirit of Trust: A Reading of Hegel's Phenomenology*. Cambridge, MA: Harvard University Press.

Brun, Georg. 2020. 'Conceptual Re-Engineering: From Explication to Reflective Equilibrium'. *Synthese* 197 (3): 925–54.

Bschir, Karim, Simon Lohse, and Hasok Chang. 2019. 'Introduction: Systematicity, the Nature of Science?'. *Synthese* 196 (3): 761–773.

Buckner, Cameron. 2023. *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artifical Intelligence*. New York: Oxford University Press.

Buckner, Cameron, and James Garson. 2019. 'Connectionism'. In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta. Fall 2019 ed.

Butlin, Patrick. 2023. 'Sharing Our Concepts with Machines'. *Erkenntnis* 88 (7): 3079–3095.

Calvillo, Jesús, Harm Brouwer, and Matthew W. Crocker. 2021. 'Semantic Systematicity in Connectionist Language Production'. *Information* 12 (8): 329.

Calvo, Paco, and John Symons, eds. 2014. *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. Cambridge, MA: MIT Press.

Camp, Elisabeth. 2004. 'The Generality Constraint and Categorial Restrictions'. *The Philosophical Quarterly* 54 (215): 209–31.

Chappell, Sophie-Grace, ed. 2015. *Intuition, Theory, and Anti-Theory in Ethics*. Oxford: Oxford University Press.

Clark, Andy. 1991. 'Systematicity, Structured Representations and Cognitive Architecture: A Reply to Fodor and Pylyshyn'. In *Connectionism and the Philosophy of Mind*. Edited by Terence Horgan and John Tienson, 198–218. Dordrecht: Springer.

Cueni, Damian, and Matthieu Queloz. 2021. 'Whence the Demand for Ethical Theory?'. *American Philosophical Quarterly* 58 (2): 135–46.

Cummins, Robert. 1996. 'Systematicity'. *The Journal of Philosophy* 93 (12): 591–614.

Cummins, Robert. 2010. *The World in the Head*. New York: Oxford University Press.

Cummins, Robert, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz. 2001. 'Systematicity and the Cognition of Structured Domains'. *Journal of Philosophy* 98 (4): 167–185.

Dancy, Jonathan. 1995. 'In Defense of Thick Concepts'. *Midwest Studies In Philosophy* 20 (1): 263–279.

Dancy, Jonathan. 2004. *Ethics without Principles*. Oxford: Clarendon Press.

Dickie, Imogen. 2010. 'The Generality of Particular Thought'. *The Philosophical Quarterly* 60 (240): 508–531.

Elgin, Catherine Z. 1983. *With Reference to Reference*. Indianapolis: Hackett.

Elgin, Catherine Z. 1996. *Considered Judgment*. Princeton: Princeton University Press.

Elgin, Catherine Z. 2017. *True Enough*. Cambridge, MA: MIT Press.

Eliasmith, Chris. 2013. *How to Build a Brain: A Neural Architecture for Biological Cognition*. New York: Oxford University Press.

Ettinger, Allyson, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. "Assessing Composition in Sentence Vector Representations." Proceedings of the 27th International Conference on Computational Linguistics.

Evans, Gareth. 1982. *The Varieties of Reference*. Edited by John McDowell. Oxford: Clarendon Press.

Fermüller, Christian G. 2010. 'Some Critical Remarks on Incompatibility Semantics'. In *The Logica Yearbook 2008*. Edited by Michal Pelis and Vit Puncochar, 81–96. Rickmansworth: College Publications.

Fodor, Jerry A. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

Fodor, Jerry A. 2003. *Hume Variations*. Oxford: Clarendon Press.

Fodor, Jerry A., and Zenon W. Pylyshyn. 1988. 'Connectionism and Cognitive Architecture: A Critical Analysis'. *Cognition* 28 (1–2): 3–71.

Fodor, Jerry, and Ernie Lepore. 2002. *The Compositionality Papers*. New York: Oxford University Press.

Fodor, Jerry, and Brian P. McLaughlin. 1990. 'Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work'. *Cognition* 35 (2): 183–205.

Frank, Stefan L., Willem F. G. Haselager, and Iris van Rooij. 2009. 'Connectionist Semantic Systematicity'. *Cognition* 110 (3): 358–379.

Franks, Paul W. 2005. *All or Nothing: Systematicity, Transcendental Arguments, and Skepticism in German Idealism*. Cambridge, MA: Harvard University Press.

Geuss, Raymond. 2020. *Who Needs a Worldview?* Cambridge, MA: Harvard University Press.

Glock, Hans-Johann. 2006. 'Concepts: Representations or Abilities?'. In *Content, Consciousness, and Perception: Essays in Contemporary Philosophy of Mind*. Edited by Ezio Di Nucci and Conor McHugh, 36–61. Cambridge: Cambridge Scholars Press.

Glock, Hans-Johann. 2009a. 'Concepts, Conceptual Schemes and Grammar'. *Philosophia* 37 (4): 653.

Glock, Hans-Johann. 2009b. 'Concepts: Where Subjectivism Goes Wrong'. *Philosophy* 84 (1): 5–29.

Glock, Hans-Johann. 2010a. 'Concepts: Between the Subjective and the Objective'. In *Mind, Method, and Morality: Essays in Honour of Anthony Kenny*. Edited by J. Cottingham and P. M. S. Hacker, 306–329. Oxford: Oxford University Press.

Glock, Hans-Johann. 2010b. 'Wittgenstein on Concepts'. In *Wittgenstein's Philosophical Investigations: A Critical Guide*. Edited by Arif Ahmed, 88–108. Cambridge: Cambridge University Press.

Glock, Hans-Johann. 2020. 'Concepts and Experience: A Non-Representationalist Perspective'. In *Concepts in Thought, Action, and Emotion: New Essays*. Edited by Christoph Demmerling and Dirk Schröder, 21–41. Abingdon: Routledge.

Goldfarb, Warren. 1997. 'Wittgenstein on Fixity of Meaning'. In *Early Analytic Philosophy: Frege, Russell, Wittgenstein*. Edited by William Walker Tait, 75–89. Chicago and La Salle, Illinois: Open Court.

Guyer, Paul. 2003. 'Kant on the Systematicity of Nature: Two Puzzles'. *History of Philosophy Quarterly* 20 (3): 277–295.

Guyer, Paul. 2005. *Kant's System of Nature and Freedom*. Oxford: Oxford University Press.

Hadley, Robert F. 1994. 'Systematicity in Connectionist Language Learning'. *Mind and Language* 9 (3): 247–72.

Hadley, Robert F. 1997. 'Cognition, Systematicity and Nomic Necessity'. *Mind and Language* 12 (2): 137–153.

Hadley, Robert F. 2004. 'On The Proper Treatment of Semantic Systematicity'. *Minds and Machines* 14 (2): 145–172.

Hare, Richard Mervyn. 1952. *The Language of Morals*. Oxford: Oxford University Press.

Hare, Richard Mervyn. 1972. *Applications of Moral Philosophy*. London: Macmillan.

Hare, Richard Mervyn. 1989. 'Ethical Theory and Utilitarianism'. In *Essays in Ethical Theory*, 212–230. Oxford: Clarendon Press.

Hare, Richard Mervyn. 2002. 'A Philosophical Autobiography'. *Utilitas* 14 (3): 269–305.

Hoyningen-Huene, Paul. 2013. *Systematicity: The Nature of Science*. New York: Oxford University Press.

Hupkes, Dieuwke, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. 'Compositionality Decomposed: How Do Neural Networks Generalise?'. *Journal of Artificial Intelligence Research* 67: 757–795.

Johnson, Kent. 2004. 'On the Systematicity of Language and Thought'. *The Journal of Philosophy* 101 (3): 111–139.

Kagan, Shelly. 1989. *The Limits of Morality*. Oxford: Oxford University Press.

Kant, Immanuel. 1929. *Critique of Pure Reason*. London: Macmillan and Co.

Keckermann, Bartholomaeus. 1600. *Systema Logicae*. Hannover: Antonius.

Kitcher, Philip. 1986. 'Projecting the Order of Nature'. In *Kant's Philosophy of Material Nature*. Edited by Robert Butts, 201–235. Boston: D. Reidel.

Kozlov, Max, and Celeste Biever. 2023. 'AI 'Breakthrough': Neural Net Has Human-Like Ability to Generalize Language'. *Nature* 623: 16–17.

Lake, Brenden M., and Marco Baroni. 2023. 'Human-Like Systematic Generalization Through a Meta-Learning Neural Network'. *Nature* 623 (7985): 115–121.

Lewis, Peter R., and Ştefan Sarkadi. 2024. 'Reflective Artificial Intelligence'. *Minds and Machines* 34 (2): 14.

MacIntyre, Alasdair C. 1978. *Against the Self-Images of the Age*. Notre Dame: University of Notre Dame Press.

MacIntyre, Alasdair C. 1988. *Whose Justice? Which Rationality?* Notre Dame: University of Notre Dame Press.

MacIntyre, Alasdair C. 2013. *After Virtue: A Study in Moral Theory*. 3 ed. London: Bloomsbury Academic.

Marchal, J. H. 1975. 'On the Concept of a System'. *Philosophy of Science* 42 (4): 448–468.

Marcus, Gary F. 2001. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.

Matthews, Robert J. 1994. 'Three-Concept Monte: Explanation, Implementation and Systematicity'. *Synthese* 101 (3): 347–363.

McLaughlin, Brian P. 1993. 'The Connectionism/Classicism Battle to Win Souls'. *Philosophical Studies* 71 (2): 163–190.

McLaughlin, Brian P. 2009. 'Systematicity Redux'. *Synthese* 170: 251–274.

Millière, Raphaël, and Cameron Buckner. 2024. 'A Philosophical Introduction to Language Models – Part II: The Way Forward'. *arXiv* arXiv:2405.03207: 1–47.

Peacocke, Christopher. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.

Perler, Dominik. 2004. 'Die Systematizität des Denkens: Zu Ockhams Theorie der mentalen Sprache'. *Philosophisches Jahrbuch* 111 (2): 291–311.

Phillips, S., and W. H. Wilson. 2016. 'Systematicity and a Categorical Theory of Cognitive Architecture: Universal Construction in Context'. *Front Psychol* 7: 1139.

Press, Ofir, M. Zhang, S. Min, L. Schmidt, N.A. Smith, and M. Lewis. 2023. 'Measuring and Narrowing the Compositionality Gap in Language Models'. *Findings of the Association for Computational Linguistics: EMNLP 2023*: 5687–5711.

Quine, Willard V. O. 1951. 'Two Dogmas of Empiricism'. *Philosophical Review* 60 (1): 20–43.

Rescher, Nicholas. 1979. *Cognitive Systematization: A Systems Theoretic Approach to a Coherentist Theory of Knowledge*. Oxford: Blackwell.

Rescher, Nicholas. 2000. *Kant and the Reach of Reason: Studies in Kant's Theory of Rational Systematization*. Cambridge: Cambridge University Press.

Rescher, Nicholas. 2005. *Cognitive Harmony: The Role of Systemic Harmony in the Constitution of Knowledge*. Pittsburgh, PA: University of Pittsburgh Press.

Rescorla, Michael. 2024. 'The Language of Thought Hypothesis'. In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta. Summer 2024 ed.

Ritschl, Otto. 1906. *System und systematische Methode in der Geschichte des wissenschaftlichen Sprachgebrauchs und der philosophischen Methodologie*. Bonn: C. Georgi.

Ross, W. D. 1930. *The Right and the Good*. Oxford: Clarendon Press.

Rudner, Richard S. 1966. *Philosophy of Social Science*. Englewood Cliffs: Prentice-Hall.

Russell, Stuart, and Peter Norvig. 2021. *Artificial Intelligence: A Modern Approach*. London: Pearson.

Ryle, Gilbert. 2009. 'Phenomenology versus 'The Concept of Mind''. In *Critical Essays: Collected Papers Volume I*, 186–204. Abingdon: Routledge.

Salje, Léa. 2019. 'Talking our Way to Systematicity'. *Philosophical Studies* 176 (10): 2563–2588.

Sellars, Wilfrid. 1958. 'Counterfactuals, Dispositions, and the Causal Modalities'. In *Minnesota Studies in the Philosophy of Science*. Edited by Herbert Feigl, Michael Scriven and Grover Maxwell. Vol. II, 225–308. Minneapolis: University of Minnesota Press.

Sellars, Wilfrid. 1997. *Empiricism and the Philosophy of Mind*. Edited by Richard Rorty. Cambridge, MA: Harvard University Press.

Shinn, Noah, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. 'Reflexion: Language agents with verbal reinforcement learning'. *Advances in Neural Information Processing Systems* 36.

Smolensky, Paul. 1988. 'The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn'. *The Southern Journal of Philosophy* 26 (S1): 137–161.

Smolensky, Paul. 1990. 'Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems'. *Artificial Intelligence* 46 (1–2): 159–216.

Smolensky, Paul, Richard Thomas McCoy, Roland Fernandez, Matthew Goldrick, and Jianfeng Gao. 2022. 'Neurocompositional Computing: From the Central Paradox of Cognition to a New Generation of AI Systems'. *AI Magazine* 43 (3): 308–322.

Spenader, Jennifer, and Reinhard Blutner. 2007. 'Compositionality and Systematicity'. In *Cognitive Foundations of Interpretation*. Edited by Gerlof Bouma, Irene Krämer and Joost Zwarts, 163–174. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Stocker, Michael. 1990. *Plural and Conflicting Values*. Oxford: Clarendon Press.

Taylor, Charles. 1985. *Philosophy and the Human Sciences: Philosophical Papers*. Vol. II. Cambridge: Cambridge University Press.

Taylor, Charles. 1989. *Sources of the Self*. Cambridge, MA: Harvard University Press.

Timpler, Clemens. 1604. *Metaphysicae Systema Methodicum*. Steinfurt: Caesar.

Travis, Charles. 2015. 'On Constraints of Generality'. *Proceedings of the Aristotelian Society* 94 (1): 165–188.

Turbanti, Giacomo. 2017. *Robert Brandom's Normative Inferentialism*. Amsterdam: John Benjamins.

van Gelder, Tim. 1990. 'Compositionality: A Connectionist Variation on a Classical Theme'. *Cognitive Science* 14 (3): 355–384.

Verdejo, Víctor M. 2015. 'The Systematicity Challenge to Anti-Representational Dynamicism'. *Synthese* 192 (3): 701–722.

Waismann, Friedrich. 1979. *Wittgenstein and the Vienna Circle*. Edited by Brian McGuinness. Oxford: Blackwell.

Williams, Bernard. 1981. 'Conflicts of Values'. In *Moral Luck*, 71–82. Cambridge: Cambridge University Press.

Williams, Bernard. 1985. *Ethics and the Limits of Philosophy*. Routledge Classics Edition. London: Routledge.

Wittgenstein, Ludwig. 1969. *On Certainty*. Edited by G. E. M. Anscombe and G. H. von Wright. Oxford: Blackwell.

Wolf, Susan. 1982. 'Moral Saints'. *The Journal of Philosophy* 79 (8): 419–439.

Wolf, Susan. 1992. 'Two Levels of Pluralism'. *Ethics* 102 (4): 785–798.

Wolf, Susan. 2007. 'Moral Psychology and the Unity of the Virtues'. *Ratio* 20 (2): 145–167.

Wolf, Susan. 2010. *Meaning in Life and Why It Matters*. Princeton: Princeton University Press.

Ypi, Lea. 2021. *The Architectonic of Reason: Purposiveness and Systematic Unity in Kant's Critique of Pure Reason*. Oxford: Oxford University Press.

Yu, Lang, and Allyson Ettinger. 2020. "Assessing Phrasal Representation and Composition in Transformers." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).