BALKAN JOURNAL OF PHILOSOPHY

Vol. 18, Issue 1, 2026

The issue is sponsored by the Bulgarian National Science Fund, competition "Bulgarian Scientific Periodicals – 2026", project No: КП-06-НП7/5 from 04.12.2025.

TABLE OF CONTENTS

Special issue devoted to the topic of "UNDERSTANDING AND KNOWLEDGE IN SCIENCE, FICTION AND ARTS"

Balkan Journal of Philosophy (2026), vol. 18, issue 1, 3 – 17

Invited paper

HOW CHATBOTS COULD UNDERSTAND LANGUAGE

Mitchell Green*

Abstract: With the recent rise of LLM technology, many researchers as well as the general public have been puzzled by the question whether the chatbots that run on that technology understand the language they produce and receive from users. Is imputing an understanding of language to chatbots legitimate, intelligible, a convenient metaphor, or just a useful fiction? Understanding is a cognitive notion, whence assessing whether something understands language turns on what kinds of psychological states we can justifiably ascribe to it. Accordingly, we here start with a delineation of the main kinds of mental state found in our own species – namely the cognitive, conative, affective, and experiential – and then argue that these four kinds need not come bundled together. That raises the question whether we might be justified in literally ascribing some such states (like cognitive and conative) to sufficiently sophisticated machines while refraining from ascribing others to them. We then provide reasons for an affirmative answer to that question, supporting the thesis that chatbots understand some of the language they produce and receive. We close by considering challenges to attitude ascription in this arena coming from (i) what we call the "attitude of one's own" objection and (ii) the embodied cognition tradition, and show they may be neutralized.

Keywords: chatbots, LLM's, attitude ascription, concepts, phenomenal concepts, understanding, inference to the best explanation.

1. Some varieties of mentality

1.1. Functionalism, and controversy over chatbot mentality

Contrary to long-discredited cases of chauvinism, one need not have a certain skin color or be of a certain gender to have a mind. So too, it is now consensus in the study of non-human animals that mentality occurs in a great variety species other than our own (Andrews, 2020; Birch et al., 2020; Newen and Starzak, 2020). Many have considered a potentially more radical step and asked whether one needs to possess neurons to have a mind. To appreciate this idea, imagine that aliens with a radically different internal composition from our own come to our planet, and not only communicate with us, but also show us how to end the Russia/Ukraine conflict, how to prevent heart disease, how to harness the sun for virtually limitless energy, and so forth. They also play chess extremely well and compose excellent music. What could justify our denying these aliens the status of having minds merely on account

*University of Connecticut, USA; mitchell.green@uconn.edu

of their lacking neurons, either centralized as in the case of mammals, or distributed as in the case of some cephalopods?

This kind of thought experiment has in turn led some to wonder if having a mind is not so much a matter of what you're made of, but of how you function. Perhaps, that is, having a mind is a matter of possessing representational states (typically but not necessarily internal) that are produced by environmental conditions, interact appropriately with one another, and then result in intelligent behavior, whether that means flourishing as in the case of a biological organism, or fulfilling a designed function as in the case of an artifact.

This suggests a general thesis of the following sort:

> *Functionalism*: An entity possessing internal representational states that are reliably brought about by what is taking place in its environment, that interact with one another in systematic ways depending on their representational content, and that tend in turn to bring about intelligent behavior in appropriate circumstances, has a mind.

The foregoing is intended to be neutral as among the several extant versions of functionalism, and only provides a sufficient condition for mentality. "Intelligent behavior" will include things like getting out of the way of oncoming traffic, graciously receiving a gift or compliment, playing a game of chess, or accurately answering questions about the Peloponnesian War. Similarly, a state's being reliably brought about by conditions in the world generally depends on its owner's having senses (or sensors) that track what is going on around it; failing that, those internal states will gather information from what is reported or otherwise communicated about the world by minded creatures or information-bearing artifacts.

Functionalism as partially described here, if further delineated in a natural way to provide necessary conditions for mentality as well, would have trouble accommodating phenomenal consciousness. However, rather than reject functionalism on this basis, we may be open to the possibility that some form of this idea could provide guidance for other aspects of mentality. Yet as we shall see presently, the phrase 'having a mind' is intolerably vague, and if we are to move forward it will require clarification of the types of mental state an entity may possess.

Historically, functionalism was embraced by pioneers of the computer revolution that began in the sixties and seventies.[1] Even though computers of the time were extremely primitive compared to what was to come, it nevertheless seemed to those innovators that computers would in time come to behave with sufficient intelligence to meet the constraints laid down by something like Functionalism as described above. This is in part because according to the functionalist approach that these pioneers embraced, intelligence is multiply realizable in the sense that, while such intelligence is in our own species instantiated in the form of a human CNS and associated body, it could in principle also be realized in the form of a sufficiently sophisticated inanimate machine.

[1] See Haugeland, 1985 for a brief historical reconstruction of this era and its sensibility. For a more in-depth history of thought about computers and mentality, see (Buckner, 2023).

A half-century or so later, computers, particularly those used by large AI firms like Anthropic, OpenAI, and the like, have powers that pioneers of the computer revolution could hardly have dreamed of. These machines have most recently been developed into "Large Language Models" that underlie chatbots like ChatGPT, Claude, and Perplexity, which have in turn impressed many of us with their ability to compete with human beings on certain tasks, or in some cases even outdo them in several others. Chatbots now show impressive performance across a wide range of tasks: They write better essays than the average undergraduate student (Herbold et al., 2023), write computer programs that compare favorably with the average software engineer (Savelka et al., 2023), rank in the 80-99th percentile on graduate admissions tests (OpenAI et al., 2023), and solve many difficult mathematical problems (Zhou et al., 2023).

In spite of the impressive achievements of chatbots, many scholars have argued that their apparent success is misleading or otherwise overblown. For instance, Bender et al. (2021) have argued that chatbots lack intentions, and are therefore incapable of meaning anything in the sense, apparently, of speaker meaning. These authors infer from this premise that the "words" that chatbots produce lack meaning (in the sense, apparently, of linguistic meaning), and conclude that chatbots are aptly compared with parrots, which make sounds but do not in fact utter words. (Hence the phrase, 'babbling stochastic parrots'.) Similarly, Fricker (2025) argues that while it may appear that chatbots are able to produce testimony (in the philosophical sense of that term), the appearance is illusory because they lack mental states, and thus lack communicative intentions.

Evidently, at least part of the controversy over the capacities of LLMs, and of the chatbots they power, is whether it is reasonable to ascribe psychological states to them. For this reason, we will focus in this essay on the question whether such devices could be said to understand any of the language they receive or produce. We will argue for a qualified affirmative answer to this question, while acknowledging that chatbots are inferior to human speakers on several dimensions.[2]

1.2. A quadripartition of psychological states

As a step toward resolving this controversy, it will help to clarify the notion of "having a mind" as adduced in Functionalism. I shall first argue that this notion admits of multiple construals due to the fact that mentality has several aspects not all of which must co-occur. In particular, human mentality may take the form of at least four types of state that are neither metaphysically nor nomologically necessitated to come in a single package. Once recognized, this fact will open the possibility of other forms of mentality exhibiting some non-empty subset of these four. Among the most prominent mental state-types in our own species are the *cognitive*, *conative*, *affective* and

[2]Green and Michel (ms) argue that chatbots may be able to perform what we may term "proto-assertions", that is utterance of indicative sentences that are not governed by all the demanding norms that govern assertion. That argument may be seen as complementing the position defended in this essay. Taken together, the two essays offer grounds for holding that chatbots are capable of some understanding of language, and may deploy that understanding in contributing to conversations by means of proto-illocutions.

experiential. Cognitive states include beliefs, (semantic) memories, degrees of belief, and states of understanding. The job of such states is to represent how things are, with each type of cognitive state doing so in distinctive ways. (We shall return to the case of understanding presently.) Conative states include plans, desires, intentions, and objectives, and play the role of guiding their possessor in efficacious action. Affective states include emotions and moods, where the latter do not mandate an object (one can be in an anxious mood, for instance from overindulging in coffee, without there being anything one is anxious about); whereas the former do (one cannot feel hopeful without there being something that one is hopeful about). While there is controversy over what if any overall role affective states play, prominent candidates include the role of revealing aspects of the world that are of concern to the agent, and motivating a proper response. (The role of fear is to reveal dangerous situations and to prompt avoidant action, etc.) (Roberts, 1988). Experiential states have a phenomenal character such as the sour taste of a lemon, or the smooth feel of velvet, and we will not here speculate on what role such states might play. Higher-order states may be constructed from any of the above four, such as when one wishes to be less jealous of a friend (conative directed toward an affective state), or knows one is smelling sulfur (cognitive directed toward an experiential state).[3]

Many cognitive and conative states have a phenomenology. When consciously entertained, my belief that I recently attended a concert might call up visual, olfactory, and auditory associations.[4] Whether those qualia are integral to the memory is not an issue we need to decide here. Instead, what matters is that cognitive states don't essentially have a phenomenal character. This is for two reasons. First, a great many of our beliefs are not consciously entertained, and while held at the unconscious level there is little reason to think that they have a phenomenal character.[5] Second, for many of our beliefs of a more abstract character, it is difficult to see what their putative phenomenal character might be: There is no particular way it is likely to believe, for instance, that there are infinitely many prime numbers, or that smoking causes cancer. An individual could of course associate imagery, and thus qualia, with

[3]We may here remain neutral on whether cognitive and other psychological states are internal to the entity to which we ascribe them, or are instead constitutively bound up with its environment. See (Varga, 2017) for further discussion of the relation between criteria for cognition and the extended mind hypothesis. Also, the quadripartition offered in the text is not meant to exhaust the types of mental state found in our species. Perceptual states, for instance, may constitute a fifth kind; on the other hand, one way to think about them is as hybrids of cognitive and experiential states, with the proviso that some perceptual states may only conceptualize their objects in minimal ways. In what follows it will not be necessary to decide on how perceptual states relate to the quadripartition offered in the text.

[4]Smithies (2013) discusses ways in which cognitive states may have phenomenal character. Also, G.E. Moore argues that even grasping a proposition has a phenomenal character, writing "it is plain that the apprehension of the meaning of one sentence with one meaning, differs in some respect from the apprehension of another sentence with a different meaning". Given what we have said about the automatic and largely unconscious character of language comprehension, we should resist accepting Moore's contention as applying generally rather than in specific cases such as those involving live metaphors. See (Dumitru, 2025) for further exploration of Moore's idea, and of the cognitive phenomenology of proof comprehension.

[5]It is consensus in contemporary cognitive science that a great many of our cognitive states are held unconsciously. See (Bermudez, 2023), especially pp. 296-301, for discussion.

one of these beliefs, but could also fail to do so and still fully grasp and accept the relevant propositions in forming these beliefs.

Similar remarks hold for conative states. Philosophers tend to focus on intention in understanding human agency and responsibility, and it is true that intending often has a phenomenology, such as when we are forming detailed plans about an upcoming course of action such as getting to the local airport for an upcoming early-morning flight. However, goals and objectives are also conative states, and these often drive actions of a more automatic character: Many of our routine daily activities are goal-directed (getting to the office at a certain time, making coffee, brushing our teeth, etc.), and can be effectively achieved with no conscious attention to the process.

Further, it is useful to note that the human mind is a product of millions of years of genetic, and hundreds of thousands of years of cultural, evolution, with many of our mental capacities adapted to the various niches in which we have survived. Had those niches been radically different, the neurotypical human mind—or at least the mind of whatever species would instead have flourished in our place – might have been correspondingly different. So too, other life forms exhibit kinds of mentality that do not easily map onto the human model. For instance, reptiles and arachnids exhibit cognitive and conative states, but we may reasonably doubt that they exhibit affective states (Birch et al, 2020).

The foregoing helps us to see that ascription of cognitive and conative states to a system does not mandate ascription of either affective or experiential states thereto. Further, an entity with a sufficiently complex system of cognitive and conative states might lack emotions and moods while competently navigating an environment in order to survive or perform tasks for which it was designed. Sensors could enable it to track what is occurring in its environment in real time without requiring it to have experiential states. These points enable us to see that we may explore the possibility of machine cognition and conation while remaining agnostic about the ability of machines to possess affective or experiential states.

1.3. Disambiguating a central question

In light of the above, it should be clear that the question, *Does this entity have a mind*? is in need of further clarification. One might be using these words to ask whether the entity has *any* of the four types of state noted above, or two, three, or all four of them. Further, unless we agree on a stipulation to the effect that having a mind necessitates having all four types of state, it will be tendentious to move from the premise that a certain entity has, say, only two of the four types of mental state noted above, to the conclusion that it doesn't really have a mind. This point is worth harping on because it's not uncommon to see scholars move from the premise that a machine lacks, say, consciousness (sometimes expressed as 'sentience'), to the conclusion that it doesn't have a mind. That inference depends on the questionable assumption that mentality of any kind depends upon the possession of one, two, or three of the other four dimensions isolated here.

2. Understanding, and understanding language

We have observed that while cognitive and conative states can have phenomenal character, they need not do so. Particularly important for our purposes is the cognitive state of understanding. Restricting discussion to understanding-that rather than understanding-how, understanding is characterized by its holistic nature, as well as its coming in degrees. For the former, it is more natural to describe someone as understanding a theory (the Krebs cycle, epidemiology) or other complex phenomenon (climate change, business etiquette in China) than it is to describe them as understanding a proposition. The former is holistic in the sense that it depends upon an appreciation of how components of a complex phenomenon interact with one another. For the coming-in-degrees component of understanding, it is natural to describe one person as having a better understanding of the Krebs cycle than does another person. Similarly, 'A knows the Krebs cycle better than B' seems just a terse way of indicating different degrees of understanding. (Hannon, 2021)

Next, while in some cases understanding has a phenomenology, in many others it does not. (I will restrict discussion to language understanding, as that is our concern in this essay.) For instance, when reinterpreting a garden-pathing sentence such as 'The horse raced past the barn fell,' in such a way as to make it intelligible, English speakers will typically have an "Aha!" experience. We may think of this as an experience of understanding. Similarly, when confronted with an image-demanding verbal metaphor in the sense described in (Green, 2017b), hearers need to form mental imagery in order to understand the utterance in which that metaphor occurs. Such imagery presumably has a phenomenal character and so requires activation of an experiential state. However, a great deal of utterances we encounter in our home language are interpreted effortlessly, and do not contain figures of speech demanding that we form imagery to make sense of them. Instead, the interpretation process will be automatic and below the threshold of conscious awareness. There need be nothing that it is like to understand a sentence such as, 'What time is it?', or 'The population of Zimbabwe has now passed 17 million people.'

What, then, is required for language understanding? Alex Barber suggests it consists in knowledge of meaning, writing

> . . . semantic theories earn their keep . . . as contributions, not to conceptual analysis of our ordinary concepts of meaning or understanding, but to an explanatory project. The phenomenon being explained will be . . . a central . . . ingredient – the semantic ingredient – of linguistic understanding. The nature of the explanation is cognitive and representationalist. Our understanding of utterances consists in knowledge of their meaning; the semantic component of this understanding is achieved through cognitive processes that generate the uttered sentence's truth conditions from internal representations (i.e. tacit knowledge) of the referential properties of its component expressions and grammar in a way that is reflected, abstractly, in the derivations within a linguistic theory. (Barber, 2013, p. 966)

Some features of Barber's proposal are worth adopting, while others are better kept at arm's length. We have noted that we wish to remain neutral on the extent to

which psychological states are internal to the subject to which they're being ascribed. For this reason, Barber's move from "representationalist" in the third sentence quoted above, to "internal representations" in the fourth should be resisted. For perhaps the representations that underlie knowledge of meaning are not internal to the system in question but are in some sense distributed within its environment. Also, recent work on the ability of LLMs to acquire language may challenge the long-held view that linguistic understanding must take the form of derivations within a linguistic theory (Piantodosi, 2024). Accordingly, we may view Barber's reference to that notion as offering only an optional way of understanding how the relevant representations operate.

On the other hand, the representations that Barber takes to enable understanding do bear explanatory weight. One way of construing these representations is as concepts associated with words or, more likely, the morphemes out of which words are made. Rather than attempt to settle any ongoing debates about the nature of concepts and their possession, let us for the moment remain neutral on the various theories of concept possession, and note that concept-possession comes in clusters: An entity that has one concept must have reasonably large number of others related it and one another in systematic ways. To have a concept of a snake one must have a concept of an animate object, and of an object that persists through time (Davidson, 1999). Notice that the concept of a snake that a non-human animal might have could be quite different from the concept that an educated adult human being has of the same animal: We might see the snake as a kind of reptile, as exothermic, as associated with biblical iniquity, and so on. A marmoset might have no such concepts and still have a concept of a snake. Or more precisely: Even though our English word 'snake' bears a meaning given in terms of concepts like exothermic, we can use that word more loosely in ascribing a concept of *snakelike creature* to the marmoset. Specialists in marmoset ecology and cognition will then be able to sharpen the concept of being a snakelike creature as it may be used to ascribe a concept to that type of monkey.

Concept possession also requires some capacity for recombination: to be able to think that a is F, one must also be able to think that a is G for some G distinct from F, and to think that b is G, for some b distinct from a. This is the well-known generality constraint on thought and associated concept possession (Camp, 2004), and it squares well with our earlier conception of understanding in holistic terms. Although philosophers tend to think of understanding an indicative sentence as grasping the proposition it expresses, and as something that a thinker can do independent of their grasp of other propositions, we now see that it takes, so to speak, a village to enable the grasp of a single proposition. If you can understand one sentence S, there must be a range of other sentences, S', S", etc., that you are also prepared to grasp without first acquiring new concepts or lexical definitions.

We may now adumbrate a conception of what it is to understand language which we may put in terms of understanding a sentence:

> *Sentence Understanding*: A system understands a sentence S just in case it possesses concepts corresponding to S's phrasal components, and combines those concepts in a way corresponding to S's grammatical structure.[6]

Although we need not build these conditions into the characterization of sentence understanding as such, we may also note two important symptoms of a system's understanding S. First is the ability of the system to discern S's truth-conditions (or satisfaction-conditions if S is not indicative); second, is the system's ability to draw appropriate inferences from other sentences to S, and from S to other sentences, that is, to appreciate S's inferential role. On this construal, sentence understanding emerges as holistic due to the facts that (i) concepts are never held piecemeal, and (ii) appreciating a sentence's inferential role involves locating it in a network of other sentences (or their semantic values).

In addition to exhibiting the holistic character of language understanding, Sentence Understanding also enables us to see how it comes in degrees. A congenitally blind person will have at best depauperate concepts of mauve and chartreuse, and a correspondingly thin understanding of sentences in which terms for these colors occur. That understanding need not be negligible, however, and the blind language user will still be able to grasp that 'mauve' and 'chartreuse' are color terms, as well as perhaps being aware of the location of the corresponding colors in the color wheel. For similar reasons, a language user lacking experiential states may have a limited understanding of terms for affective states with strong experiential character, such as 'grief' and 'terror' but may still have enough conceptual scaffolding associated with these terms to engage in limited conversation involving them. This is presumably what designers of the recent wave of therapy bots are hoping to achieve (Heintz et al., 2025).

3. Ascribing psychological states

We have so far explored what is involved in understanding language, including its construal as a holistic cognitive state coming in degrees. Given the independence of cognitive states from the other three types of state explored above, we should expect that an entity can understand language in spite of being bereft of phenomenal consciousness or affect. This leaves open the possibility that certain concepts – the so-called phenomenal concepts – can only be fully grasped by those capable of experiential states. If that possibility is realized, then given our agnosticism about the ability of inanimate objects to have experiential states, that will only show that inanimate objects may be unable fully to understand certain concepts and the terms that express them. It will not show that they are able to understand language across the board.

However, nothing we have established so far implies that inanimate objects like chatbots understand language. What sorts of considerations would justify such a claim? To answer this question, we begin with some methodology from the philosophy of science.

[6] For a different formulation of the nature of linguistic understanding, see (Gurova, 2025).

3.1. Inference to the Best Explanation as a framework for attitude ascription

A methodological principle guiding much theory choice is inference to the best explanation, according to which we are justified in accepting that theory that best explains a range of phenomena that is of interest. More precisely,

> *Inference to the Best Explanation* (IBE): If from among a set T_1, ..., T_n of theories, each of which explains phenomenon D, T_n is the best, then we are justified in accepting T_n. (Lipton, 2000).

IBE forces us to confront subtle questions about what makes one theory better than another. This is not the place to pursue such questions in detail, but it is not controversial to contend that among the virtues helping to make one theory better than another, we will find simplicity, coherence with established theories, and fecundity, in the sense of a theory's giving rise to novel predictions that end up being confirmed. Related to simplicity is the notion of intelligibility: If a theory is so complex as to be unintelligible to even the most powerful human minds, that is a strike against it. So too, simpler theories are *ceteris paribus* more parsimonious than their rivals. A theory that posits fewer theoretical entities than its rivals, or fewer axioms, will to that extent be superior to them. If two theories are tied for the status of 'best', then the wisest choice may be to remain agnostic between them until further evidence is obtained favoring one over the other. In some cases, seemingly distinct theories may be notational variants of one another, in which event it may appear that the choice between them should be made, if at all, on pragmatic grounds.[7]

Applied to the case at hand, IBE dictates that if the best explanation of an entity's behavior involves imputation of psychological states, then we are justified in ascribing such states to that entity. For instance, most of us would agree that our gross bodily actions like reaching for a glass of water, or pressing on a bike's pedals, are brought about by the concerted actions of huge numbers of neurons and muscle cells. However, few would infer from this that folk-psychological explanations of behavior are for this reason illegitimate. The reason is that folk psychological explanations are on the whole superior to those given in neurological terms, not least because the former are intelligible while the latter are generally not.

3.2. The redescription fallacy

What's good for the animate goose should be good for the inanimate gander. We do well not to employ a double-standard when considering inanimate objects. After all, it is often remarked that LLMs are "really" just next-token predictors that have been trained via exposure to vast amounts of human-generated text, and then it is inferred on this basis that no higher-level ascriptions are justified.[8] However, as Milliere and

[7] Green (1999), for instance, argues that there may be no fact of the matter whether belief-desire psychology, or instead a decision-theoretic approach on which we ascribe degrees of belief and preference orderings, is correct.

[8] Butlin (2023) for instance argues that because LLMs are designed to predict a next token after in an input string, they are incapable of representing any of the worldly entities that those tokens might denote, writing, "GPT-3 therefore appears to be capable of representing the words 'skin', 'red' and 'rash'

Buckner (2024) observe, this line of reasoning is fallacious. For them, the suspect line of reasoning …

> …arises when critics argue that a system cannot model a particular cognitive capacity, simply because its operations can be explained in less abstract and more deflationary terms. In the present context, the fallacy manifests in claims that LLMs could not possibly be good models of some cognitive capacity ϕ because their operations merely consist in a collection of statistical calculations, or linear algebra operations, or next-token predictions. Such arguments are only valid if accompanied by evidence demonstrating that a system, defined in these terms, is inherently incapable of implementing ϕ. To illustrate, consider the flawed logic in asserting that …brain activity could not possibly implement cognition because it can be described as a collection of neural firings. The critical question is not whether the operations of an LLM can be simplistically described in non-mental terms, but whether these operations, when appropriately organized, can implement the same processes or algorithms as the mind, when described at an appropriate level of computational abstraction (p. 10).

As these authors point out, it is fallacious to move from the premise that a complex system's behavior can be exhaustively described in lower-level terms, to the conclusion that no higher-level descriptions are applicable. Let us call this the *redescription fallacy*. Bearing in mind the difference between description and explanation, we may readily observe that such a complete description need not provide an explanation of the phenomenon we are hoping to explain. That is, the redescription fallacy is indeed a fallacy.

That said, some aspects of Milliere's and Buckner's remarks merit refinement. First of all, the central question is not whether LLM computations can implement the same processes as the human mind. After all, mentality has at least four distinct dimensions, and it is not fruitful to assume that the only way for a machine to exhibit mentality is for it to do so in the same way the human mind does. As we have seen, it may be that machines exhibit mentality in ways quite different from what we encounter in our own species.

Second, Milliere and Buckner's point needs qualification in terms of IBE: Even if higher-level ascriptions to a machine might enable us to see it as implementing some of the same processes as does the human mind, those ascriptions may not be part of the best explanation of what that machine does. The question is whether the ascription of mental states is doing some explanatory work that a more parsimonious theory cannot do.

With these refinements in mind, it should now be clear that the contention that LLM's merely predict the next token to follow in a sequence is tendentious. For while it is true that they do make such predictions, it does not follow that higher-level ascriptions are inapplicable. As elsewhere, the issue is whether such ascriptions do any explanatory work that is not done by the lower-level description.

---

but not the worldly entities to which they refer to" (p. 3084). What is being missed here is the possibility that a more semantic description of the LLM might bear explanatory weight.

3.3. Ascribing attitudes to chatbots

This is where Generative AI has a role to play. The reason is that Generative AI is now producing outputs whose provenance we can only understand in a schematic way in the absence of attitude ascription. When a chatbot performs well on a standardized test, programs code, composes poetry, or solves a problem in mathematics, its ability to do so is not well explained by its having been trained to do next-token prediction. After all, the poem it produces, for instance, has never been written before. Instead, when I ask a chatbot to, for instance, write a limerick about surfing in a bathtub, it produces five lines of novel text according to the rhyme scheme of the genre, on the topic of surfing in a bathtub that is at least marginally witty. We may account for this by supposing that the chatbot understands the task given it, adopts the goal of fulfilling that task, and then produces the requested limerick in the belief that it will meet the user's request and thereby fulfill the chatbot's goal.

The chatbot's ability to fulfill this request depends on its possession of concepts enabling it to understand the language that it receives and produces. That does not mean that the concepts in question map straightforwardly onto human concepts. As Milliere and Buckner (Ibid) point out, we do well to think of the concepts ascribable to LLMs as vectors in an n-dimensional space (for potentially very large n). On this approach, two words are construed as similar in meaning if they are close to one another in that space; disparate in meaning if they're far apart. Because mathematical operations (addition, subtraction, multiplication, etc.) can be performed on vectors, such operations may also be performed on those vectors representing concepts. Thus for instance, by subtracting the vector for 'man' from the vector for 'king' and adding the vector for 'woman,' we can obtain a vector that is close to the vector for 'queen'.

In this light, we may now see that there is good reason to think that the best explanation of an LLM's use of language is that it was moved by two kinds of psychological state: beliefs (cognitive) and goals (conative), each of which is underpinned by concepts. Unlike the remark that the machine is a next-token predictor, ascribing a pair of such states helps to explain why the machine does what it does in a particular case. If we are at pains to avoid chauvinism about skin color, gender, and species, we should also be prepared to evaluate chatbots and the LLMs underpinning them on their own terms, and take seriously the possibility that they exhibit some aspects of mentality.

4. Challenges to Chatbot Understanding

We have now provided reasons for holding that sophisticated chatbots understand the language they receive and produce, *modulo* the qualification that their understanding of language involving phenomenal concepts will be compromised as compared to how it is grasped by neurotypical adult human speakers. So too, in light of our agnosticism about whether these machines can experience affective states, we are in no position to suppose that they could empathize with their interlocutors. At least on a construal of (affective) empathy as involving imagining one's way into another's affective situation, a machine lacking experiential or affective states would appear unable

to do such a thing.[9] Yet as with the case of language involving phenomenal concepts, the most this shows is that chatbots are inferior to other interlocutors, not that they are incapable of understanding language or engaging in useful verbal interactions.

In this final section it will be helpful to consider some objections to the possibility of chatbots understanding language.

4.1. Objection 1: "Reasons of one's own"

Some authors suggest that machines do not have psychological states of their own, and infer from this that they do not possess such states at all. For instance, Butlin and Viebahn (2023) contend that AI systems do not have interests of their own, and apparently infer from this premise that they do not have interests. Instead, it would appear that any interests that such systems *seem* to have are really just those of their designers. Let us assume that a psychological state such as a belief or objective cannot belong to two agents simultaneously. Then, if state Ψ belongs to agent 1, then it cannot also belong to any distinct agent.

Butlin and Viebahn are responding to an argument of Green and Michel (2022) that included the contention that humanoid robots could within a generation start to police our streets and highways, re-directing traffic after accidents or in dangerous weather situations, ticket speeders, and so on. Such "Robot Cops", these authors contend, could in the course of serving their human police masters, perform speech acts. Leaving aside the merits of this argument, it may here be noted that as imagined, the Robot Cops have sufficient autonomy from the humans who deploy them that these humans may be unaware of much of what they are doing. As such, many of the psychological states that we may wish to ascribe to a Robot Cop are not had by any human being. Instead, at most, such states are causally traceable to decisions made by their human bosses. However, it by no means follows that those states are not those Cops' own.

Similarly, one may be tempted to suggest that chatbots lack psychological states of their own, on account of the fact that they are trained by exposure to huge amounts of human-generated text as part of their training regimen. Here too, however, the great majority of psychological states we may be tempted to ascribe to a chatbot are not had by any human being, living or otherwise. It is highly unlikely, for instance, that any human being came up with the surfing-in-the-bathtub limerick prior to my soliciting it from the chatbot. So the thought that underlies the limerick does not belong to anyone else. While that thought may not be terribly original or deep, that possibility is not relevant to whether a chatbot can have it.

4.2. Objection 2: embodied cognition

In their contribution to the popular periodical *The Conversation*, 'It takes a body to understand the world: Why ChatGPT and other language AI's don't know that they're

[9]Green (2010b) defends a conception of affective empathy as a matter of imagining one's way into another's affective situation.

saying,' (Glenberg and Jones, 2023), the authors report that chatbots are not adept at answering questions that would be fairly easy for a 10 year old child. (This article crystallizes research described in more detail in Glenberg and Jones (2023), and Jones et al. (2022).) For instance, when asked which would be better for use to fan coals for a barbecue, a paper map or a stone, GPT-3 offered the latter. The same technology chose a hairpin over a warm thermos when asked which would be more effective in smoothing a wrinkled shirt. Again, when queried as to which artifact would serve best to cover one's hair for work in a fast-food restaurant, a paper sandwich wrapper or a hamburger bun, the chatbot went for the bun.

The authors reporting these results diagnose the chatbots' mistakes as due to their not being embodied. The thought is that it is only by virtue of having a body that enables one to discern an object's affordances: Having hands and arms enables one to appreciate the fact that a paper map will work better on the coals than on a stone – at least if the stone is of a typical shape as far as stones go; and having a head enables one to discern that the sandwich wrapper will do better in keeping hair from falling from it than the hamburger bun.

All this is reasonable enough, and we do well to reflect on these points in considering the extent to which human understanding of language rests on our nature as embodied beings. But three points should help us to see that the issue raised by these authors is not fatal to the line of thought defended in this essay. First, it should come as no surprise that a chatbot will be inferior in its mastery of certain areas of language, as well as of commonsense knowledge, as compared to normal human users. We are already familiar with how chatbots are likely to have a limited understanding of language expressing phenomenal concepts, as well as of concepts that would enable them to empathize with the emotions experienced by their human interlocutors. This does not show that chatbots fail to understand language.

Second, the authors contend that "The meaning of a word is intimately related to the human body". As an example supporting this claim, the authors give, 'paper sandwich wrapper', the understanding of which includes not just what it is used for, but also how it feels when squeezed into a ball, and how far it can be thrown. Our question should be how far this point generalizes. Is the meaning of terms like 'therefore', 'and', and 'most' intimately related to the human body? If so, how? After all, even if my understanding of 'and' carries an association with adding things to a heap, it's not clear how that association affects my understanding of language in which the word occurs. These points hold as well with terms for abstract concepts like 'justice' and 'heterozygote'.

Third, while Glenberg and Jones begin their article by diagnosing the problem as being due to chatbots not understanding language the way humans do, they conclude that same article with a much stronger claim:

> ChatGPT is a fascinating tool that will undoubtedly be used for good – and not-so-good – purposes. But don't be fooled into thinking that it understands the words it spews, let alone that it's sentient (Ibid).

The authors thereby move from the premise that chatbots don't understand language the way humans do, to the conclusion that they don't understand language. This is a non sequitur, and for it to be a good inference we would need a defense of the further premise that the human way is the only possible way of understanding language. I hope in these pages to have given reasons for doubting that premise.

References

Andrews, K. (2020). *How to study animal minds*. Cambridge University Press.

Barber, A. (2013). Understanding and knowledge of meaning. *Philosophy Compass, 8*, 964–977.

Bender, E., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–622). ACM. https://doi.org/10.1145/3442188.3445922

Bermúdez, J. (2013). *Cognitive science: An introduction to the science of the mind* (4th ed.). Cambridge University Press.

Birch, J., Schnell, A., & Clayton, N. (2020). Dimensions of animal consciousness. *Trends in Cognitive Sciences, 24*, 789–801.

Buckner, C. (2013). *From deep learning to rational machines: What the history of philosophy can teach us about the future of artificial intelligence*. Oxford University Press.

Butlin, P. (2023). Sharing our concepts with machines. *Erkenntnis, 88*, 3079–3085.

Butlin, P., & Viebahn, E. (2025). AI assertion. *Ergo*. https://doi.org/10.3998/ergo.7960

Camp, E. (2004). The generality constraint and categorial restrictions. *The Philosophical Quarterly, 54*, 209–231.

Davidson, D. (1999). The emergence of thought. *Erkenntnis, 51*, 7–17.

Dumitru, M. (2025). "Feeling the proof": Is there such a thing as the phenomenology of reasoning? In A. Mărășoiu & M. Dumitru (Eds.), *Understanding and conscious experiences: Philosophical and scientific perspectives* (pp. 179–192). Routledge.

Fricker, E. (2025). On the metaphysical and epistemic contrasts between human and AI testimony. *Inquiry*. https://doi.org/10.1080/0020174X.2025.2553297

Glenberg, A., & Jones, C. R. (2023). It takes a body to understand the world – Why ChatGPT and other language AIs don't understand what they're saying. *The Conversation*. https://doi.org/10.64628/AAI.w6mhj6hkx

Green, M. (2017). Imagery, expression, and metaphor. *Philosophical Studies, 174*, 33–46.

Green, M. (2010a). Language understanding and knowledge of meaning. *Baltic International Yearbook of Language, Logic and Communication, 5*, 1–17.

Green, M. (2010b). How and what can we learn from literature? In G. Hagberg & W. Jost (Eds.), *The Blackwell companion to the philosophy of literature* (pp. 350–366). Wiley-Blackwell.

Green, M. (2025). Large language models and the varieties of meaning. *Philosophy of AI, 1*, 34–40.

Green, M., & Michel, J. (2022). What might machines mean? *Minds and Machines, 32*, 323–338.

Green, M., & Michel, J. (manuscript under review). *What could a chatbot mean?*

Gurova, L. (2025). Understanding and inference in recent works on scientific understanding. In A. Mărășoiu & M. Dumitru (Eds.), *Understanding and conscious experiences: Philosophical and scientific perspectives* (pp. 62–74). Routledge.

Hannon, M. (2021). Recent work in the epistemology of understanding. *American Philosophical Quarterly, 58*, 269–290.

Haugeland, J. (1985). *Artificial intelligence: The very idea*. MIT Press.

Heintz, M., et al. (2025). Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI, 2*. https://doi.org/10.1056/AIoa2400802

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports, 13*(1), 18617.

Jones, C., Chang, T., & Coulson, S. (2022). Distributional semantics still can't account for affordances. *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*(44).

Lipton, P. (2000). Inference to the best explanation. In W. Newton-Smith (Ed.), *A companion to the philosophy of science* (pp. 184–193). Blackwell.

Millière, R. (forthcoming). Language models as models of language. In R. Nefdt, G. Dupre, & K. Stanton (Eds.), *The Oxford handbook of the philosophy of linguistics*. Oxford University Press.

Millière, R., & Buckner, C. (2024). *A philosophical introduction to language models, part I: Continuity with classic debates*. arXiv. https://arxiv.org/pdf/2401.03910

Mitchell, M., & Krakauer, D. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, 120*. https://doi.org/10.1073/pnas.2215907120

Moore, G. E. (2014). Propositions. In *Some main problems of philosophy* (pp. 52–71). Routledge. (Original work published 1953)

Newen, A., & Starzak, T. (2020). How to ascribe beliefs to animals. *Mind & Language, 37*, 3–21.

OpenAI. (2023). *GPT-4 technical report*.

Piantadosi, S. T. (2023). Modern language models refute Chomsky's approach to language. In E. Gibson & M. Poliak (Eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett* (pp. 353–414). Language Science Press.

Roberts, R. (1988). What an emotion is: A sketch. *The Philosophical Review, 97*, 183–209.

Sambrotta, M. (2025). LLMs and the logical space of reasons. *Minds and Machines, 35*, Article 46. https://doi.org/10.1007/s11023-025-09751-y

Savelka, J., Agarwal, A., An, M., Bogart, C., & Sakr, M. (2023). Thrilled by your progress! Large language models (GPT-4) no longer struggle to pass assessments in higher education programming courses. In *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1* (pp. 78–92).

Smithies, D. (2013). The nature of cognitive phenomenology. *Philosophy Compass, 8*, 744–754.

Stoljar, D., & Zhang, Z. (2024). Why ChatGPT3 doesn't think: An argument from rationality. *Inquiry*. https://doi.org/10.1080/0020174X.2024.2427061

Varga, S. (2017). Demarcating the realm of cognition. *Journal for the General Philosophy of Science, 49*, 435–450.

Zhou, A., Wang, K., Lu, Z., Shi, W., Luo, S., Qin, Z., Lu, S., Jia, A., Song, L., Zhan, M., & Li, H. (2023). *Solving challenging math word problems using GPT-4 Code Interpreter with code-based self-verification*. arXiv. https://arxiv.org/abs/2308.07921

Balkan Journal of Philosophy (2026), vol. 18, issue 1, 18 – 31

UNDERSTANDING THE SURREALIST EYE. THE EPISTEMOLOGICAL SHIFT IN PICTORIAL REPRESENTATION

Constantin Stoenescu*

Abstract: In his famous *Manifeste du Surréalisme* André Breton proposes the exercise of creative freedom through detachment from the objective external world. But what prevents us from playfully freeing ourselves from the external world, from the first-instance reality, in order to pass into super-reality? The organ that keeps us fixed in the captivity of so-called objectivity, in the so-called "iron cage" of a rationality that wants to capture the world independent of our minds, is the eye. Therefore, in order to pass from looking outward, from the perception that puts us in contact with the object, to looking inward, to the introspective plunge, the biological eye must be annihilated in favor of the inner eye, not only introspective, but also oneiric and occult. This aesthetic program is based epistemologically on a critique of representational realism and the rejection of the so-called "myth of the given", the effect of belief in an "innocent eye". My goal in this paper is to identify the epistemological assumptions that underpinned traditional theories in art, primarily pictorial representational art, and to identify the change at this level that led to new trends in 20th-century art, especially those that made the surrealist movement possible. Considering Victor Brauner as a relevant case, I propose to understand the surrealist eye in three artistically and aesthetically relevant ways, namely, the "suppressed eye", the "autonomous eye" and the "sublimated eye".

Keywords: representational realism, *mimesis*, "myth of the given", "the innocent eye", conventions and perception, surrealist eye.

1. The tradition of representational realism

1.1. The ancient Greeks' insight: truth and mimesis

The idea of art as representing or imitating or copying nature or reality comes from the ancient Greeks and it was expressed through the concept of *mimesis*.[1] I shall try to propose a short reconstruction of the ancient Greek approach to representation, taking into account the works of Plato and Aristotle. Although Plato mainly focuses on poetry, some of his statements are applicable to the case of pictorial representation.

The debate opened by Plato in *The Republic Book 3* about the role of poetry in the ideal state is well known especially as the first defence of censorship. The

[1] I have in mind those contextual uses of the term "mimesis" in which it should be understood as "representation" or "copy", but I draw attention that this translation does not exhaustively capture the meanings in which the term was used by the ancient Greeks. For example, "mimesis" was used sometimes, as in the case of theatrical performances, with the meanings of "mimicry" or "enactment".

*Faculty of Philosophy, University of Bucharest; constantin.stoenescu@filosofie.unibuc.ro

core of Plato's argument is that poetry conveys misrepresentations, which is why he establishes both content and formal restrictions. Plato claimed that the poets make "stupid mistakes" (Plato, *The Republic*, p. 379), or that some stories are "quite untrue" (Plato, *The Republic*, p. 387). But in *Book 10* Plato's argument becomes more radical and it takes the logical form of a universal proposition: "We may assume (...) that all the poets from Homer downwards have no grasp of reality, but merely give us a superficial representation of any subject they treat including human goodness" (Plato, *The Republic*, 598-600). The poetry (and all the other arts), according to Plato's view in his middle period, when *The Republic* was written, is similar with the beliefs derived from senses, while true knowledge is of the Forms and is acquired by reason. Through the senses we have access only to appearances.

Generally, the so-called "art of representation" is compared by Plato to holding a mirror in your hand, which anyone can do, even a fool, meaning you wouldn't need to have any knowledge or skill to hold a mirror. But he then admits that a painter needs some skills, like any craftsman, and concludes that a skill involves some knowledge, but the content of this knowledge would be only one of mere appearances, similar to the reflections in a mirror. A painter guided by Plato's doctrine should try to represent only appearances, not the real and true world of the intelligible Forms.

But among painters in Plato's time there was a debate about the differences between representing an object like a chair and representing the beauty of a chair, or between painting a person who was supposed to resemble the real person and painting a person who has a certain character, in which case the character of someone cunning, to give an example, is more important than the person as such. It also seems that the painters did not compose a character according to a single model, but were inspired by different models, taking the eyes from one, the nose from another, and so on.

From a metaphysical perspective, Aristotle agreed with Plato that the same name is given to particular different things if they have something in common, but he rejected the idea that this common feature is their resemblance to an intelligible Form. Aristotle's view, described at length in *The Poetics* was that the so-called universal Forms aren't transcendent, but they exist in our sensible world as essences of the real objects perceived by us. Aristotle also agreed with Plato that the arts are forms of imitation or *mimesis*, but he introduced in his Poetics the idea that there are different cases of imitation because different arts use different media, such as colours, shapes, sounds or words and they do this in different ways.

1.2. The epistemological weakness of mimesis

I think it should be noted that this debate about *mimesis* in ancient Greek philosophy is conducted rather in terms of a concept of truth understood as a correspondence between language and reality, and only sometimes does it seem to be moved into the context of a theory about the representation of the external world as a real one. This resulted in a certain ambiguity that over time led to the crediting of the idea that, according to Aristotle, art must imitate nature. Hence the persistent idea that pictorial

art imitates nature (or reality) through its mimetic copying, that is, it has to mirror or re-present it as it is.

From here it was only a step to the idea that an artistic representation could be so truthful or veridical as to be taken as real, in other words, the eye could be deceived if a representation seemed to be real. Such a result is not at all consistent with the assumptions of the Aristotelian theory of truth. We may talk in such a case about an image with no correspondence in reality, but only in the mind of the viewer who perceives it, and who subjectively represents it and is aware of it. If the pictorial image is veridical, namely, it resembles reality, it can be considered to be a realistic representation because it looks like a real object, but it really has no counterpart in the external world, and from this perspective it cannot be discussed in terms of truth.

Therefore, from an epistemological point of view, the theory of *mimesis* as a copy is compatible with a representational theory of perception, even a realistic one, which leaves room for the validity of consequences that seem to contradict the initial premises. The idea is that the image can seem real even though it has no counterpart in reality, and it is in this sense an illusion. Epistemologically acceptable, such a development of the theory of *mimetic* representation leads to some inconvenient ontological consequences, such as the problem of the ontological status of these representations through images that seem real but have no counterpart in reality.[2]

However, regardless of these metaphysical tribulations, the idea remains that the eye can be deceived into taking as reality what is merely a representation of reality. I would mention the stories about paintings that deceived not only people but also animals. Pliny the Elder (See Pliny the Elder, pp. 310-312) told a story about two rival painters, Zeuxis and Parrhasius, who competed in 397 B.C. to see whose painting was the most realistic. Zeuxis had painted some grapes so realistically that birds tried to peck the fruits, while Parrhasius painted a curtain and Zeuxis himself was deceived because he tried to lift the curtain to see the picture, but the picture was the curtain itself–a picture of a curtain.

Generally speaking, the eye is deceived because it takes as real a representation of something which is not real, but only an illusion. From an epistemological point of view, representational realism remains coherent if we admit the possibility of illusions and hallucinations; moreover, I think that the production of illusions and hallucinations is an argument for the representational realism as a theory of perception in competition with the so called theory of direct realism. Therefore, illusions are possible and the eye may be deceived. Even we accept the theory of *mimesis* framework with all its constraints.

[2] I would venture to say that Plato intuited such consequences and, as a result, in his late dialogues, in *Timaeus* and *Politicus* (*Statesman*), he tried to develop an alternative to the theory of mimesis proposed in the previous dialogues of his youth and middle period, advancing the hypothesis of a mimesis without the mediation of the image. However, there is neither time nor place here to explore this aspect in detail.

1.3. Representation and the act-object theory of perception

My view is that far from being dogmatically closed to a theory of *mimetic* representation of reality, the thought of the ancient Greeks contains the germinal seeds of a representational approach proper to an act-object theory of perception for which the possibility of illusion is rather a confirmation than a bewildering difficulty of it. In other words, the possibility of illusions and hallucinations validates one of the major consequences of an act-object theory, namely, that the epistemic subject is not in a direct representational contact with objects of the external world, but through the mediation of the re-presentations of these objects that are given to him perceptually and of which the subject is aware. Even if this mediation leads from an epistemological point of view to the difficulties specific to a "veil of perception" between the subject and the world, and from an ontological point of view to a multiplication of entities by introducing these intermediate mental objects called "representations" whose status must be elucidated, the act-object theory inferentially guarantees the best explanations regarding the process of pictorial representation of the external world.

Let's enter the details of an analysis of the pictorial representational illusion. An interesting sub-class of misleading or deceiving images is that of *trompe l'oeil* paintings. They may be considered representational pictures because they stand in for a real object, they look like a real object, but at the same time, they aren't copies of reality because there is no real object to be copied by them. They are just illusory representations that replace the real object and they are taken as real even if they are just appearances. We have knowledge of the corresponding real object, for example, we know what a bunch of grapes or a curtain looks like, and we judge based on the perceptual information we have. All the blame falls on the senses that deceive us.

On the other hand, we know from the history of painting that beings and entities such as Adam and Eve, Pegasus or "Knowledge" have been represented, although we either do not know how they look in reality–in the case of Adam and Eve–if we may speak about reality, or we only imagine them, as in the case of Pegasus, or we know from the very beginning that the representation is a cultural convention, as in the case of "Knowledge". This makes us able to recognize a painting in which Adam and Eve taste the apple from the tree of knowledge, that is, we know that they are Adam and Eve and no one else. We also recognize Pegasus because it is as we imagine it, although we know that such a being has not been seen in reality, and, more than that, we recognize "Knowledge" in the fight with "Ignorance" because we know how each of them could be anthropomorphized and we imagine the two characters in this way, according to our shared cultural conventions.

As a result, when we say that a painting is a correct representation, we must distinguish between the case in which we consider the mimetic character of a representation, in the sense of a copy of a real object, and the case in which correctness presupposes a conscious complicity of the eye, in the sense that it allows itself to be deceived by one image because it is accepted from the perspective of certain expectations or cultural conventions or other consensually accepted beliefs. But this already means accepting that the eye is not just a passive receptor and does not only have the

function of representation as a mirror of reality. In fact, we now realize this truth–no one has ever worked with such an eye in an absolute sense. Let us remember that Plato admitted as early as *The Republic*, when he was developing the theory of participation, a double ambiguity of the image towards both sensible things and intelligible Ideas.

2. "The innocent eye" and its epistemological presuppositions

2.1. The myth of the given

Both the idea that an image must be a copy of reality so that the copy can stand in place of the real object and the idea that such a perfect copy can deceive the eye are derived from a radical and erroneous interpretation of the theory of *mimesis*, an interpretation that loses along the way the very specificity of the thesis about *mimesis*. The copy is an imitation of reality and, being a copy, precisely for that reason, no matter how perfect, it is not the same as the real object but remains within a representational horizon.

This radical interpretation of the idea of *mimesis*, somewhat naive, perhaps precisely for that reason attractive and resilient, is based on what Ernst Gombrich described as "the myth of the innocent eye."[3] This myth is supported by the assumption that the act of seeing something is direct, in the sense that the eye passively perceives, like a perfect mirror, what is given to us, without any distortion in the representational process. Otherwise, we can view a picture without preconceptions.

The British empiricists developed this approach–Locke systematized it, and Hume took it to its ultimate consequences. In his *Essay Concerning Human Understanding*, Locke rejected the Cartesian theory of innate ideas and concluded that initially our mind is like a "white paper, void of all characters, without any ideas", and that its content is furnished by experience, that "our observation employed either about external sensible objects, or about the internal operations of our mind, perceived and reflected on by ourselves, is that supplies our understandings with all materials of thinking" (Locke, Essay, Book II, Chapter I, paragraph 2). Hume refined Locke's empiricist thesis with the help of the distinction between impressions and ideas and concluded that "all our simple ideas in their first appearance are derived from simple impressions, which are correspondent to them, and which they exactly represent"; but, at the same time, he admitted that "though there is, in general, a great resemblance, betwixt our *complex* impressions and ideas, yet the rule is not universally true…" (Hume, 1982, pp. 47-48).

But the main epistemological temptation common to all these theories of perception from the ancient Greeks to British empiricism is the view that knowledge (by acquaintance) of what we perceive could be independent of the conceptual processes embedded in perception. This is the sense-data theory of knowledge, related with all the foundational approaches, and it is based, using an expression proposed by Wilfrid

[3]This expression is used largely by Gombrich in his *Art and illusion* (1977).

Sellars (1956), on the so-called "myth of the given". Sellars correlates the traditional representational theory of sense-data with the so-called "*Fido*" - *Fido semantic principle*[4] by which sensory contents get a linguistic description which leads to the idea that they can be described neutrally, independently of any previous knowledge, other than knowledge of language as an ability, namely, as a *knowing how*. According to Sellars interpretation of conceptual empiricism, the initial elements of any experience are, for example, the impressions of a certain red thing and not the impression of red as such. As a result, we become aware of the repeatability of impressions based on association between words (e.g. "red") and classes of similar things. But in this case, we will have to clarify what it means that "x resembles y" and if the mediation is exclusively a linguistic one. In this way, we arrive at a psychological nominalism by which we equate concepts with words, and thoughts with verbal acts, as if thinking were just a problem of language acquisition and use. The fact that we have the sensation of red does not presuppose that the abstract idea of red (or "redness" or "Red") exists in our mind, but only that we have the natural ability to perceive red things. Therefore, on this basis we will use the word "red" to name the sensorial quality of red and we will say that we understand the word "red". In other words, we can learn concepts with the help of sensory stimuli and under the conditions of the correct use of language according to the rules of meaning: "'Red' means a red quality" (Sellars, 1956, pp. 289-290).

But this connection between sensory data and language ultimately leads to the conclusion that the description of sensory contents depends on language, and a description is theoretically loaded, namely, it is full of the tacit knowledge-ingredients incorporated into the language and used by the community of observers. This epistemological idea had already been proffered by Karl R. Popper who warned that his term "empirical basis" has ironical overtones, suggesting that the basis is not firm, and that any observational statement, like all languages, "are impregnated with theories" (Popper, 2002, pp. 93-94). Even a so-called phenomenal language, let's suppose, one about colours, is impregnated by theories about space, time and colours. Moreover, another philosopher, Michael Polanyi (1958) emphasizes the role of the tacit dimension at the level of personal knowledge available to the epistemic subject, and N. R. Hanson (1958) uses the distinction between "seeing as" and "seeing that", arguing that what we see is always filtered through our existing information and preconceptions.

All these epistemological ideas, as Sheldon Saul Richmond (1994) argued, are the roots for a new critical understanding of all the new trends in the arts, which is developed by Ernst H. Gombrich in his *Art and Illusio*. Gombrich, as he himself admits in the preface to the first edition of his book, applied many epistemological ideas, especially taken from Karl R. Popper, to the story of visual discoveries in art. Knowing and seeing are related by Gombrich both with the tradition to which an artist belongs and with the pre-existing "schemata", some of them in the form of conventions, that are learned and used in the interpretation of sensory data and in the process of making and matching and remaking the final and finished pictorial image.

[4]This label was proposed by Gilbert Ryle in his "The Theory of Meaning" (1962).

2.2. A deconstruction of the innocent eye

To enter into the details of Gombrich's theory about perception and the visual arts, I follow Rosalind Hursthouse's proposal of deconstructing the myth of the innocent eye into three assertions, each of which can be neutralized and falsified, thus rejecting the entire argument.

First, the expository enumeration of the constitutive premises (See Hursthouse, 1992, p. 265):

1. We are able to consciously make the distinction between *the reception* of visual data or information and *the interpretation* of it, namely, between *seeing* and *seeing as*.

2. Reception comes first, interpretation follows it. Interpretation depends on reception and the reception is prior and independent of any interpretation. The data are given directly, without any interpretative filter or processing.

3. Reception is the same for all epistemic subjects, if perception occurs under standard conditions and the subjects are cognitively and consciously normal, while interpretation may or may not vary from one person to another depending on other factors such as one's own previous perceptual experiences, perceptual memory, various expectations or the cultural background associated with the visual experience.

Each of all these three premises may be easily denied by their corresponding contrary claims:

1. It is difficult to distinguish between the reception of visual data and their interpretation. We are rarely conscious of this sort of differences. Visual experience is a flow based on the permanent activation of memory to give identity to perceptions, otherwise we would live in a continuous present in which each experience would be a new experience, without noticing similarities to previous experiences of the same type, which means that, as a rule, we do not just see, but we see *as*. Moreover, even an unconscious visual perception has an inferential character in the sense that what we see is implicitly processed by our system of beliefs and, as a rule, in the case of such perceptions that have become stereotyped during our psychic development, we only become conscious of the case when our expectations are not met.

2. The perceptual act depends on interpretation as much as interpretation depends on the reception of perceptual data. In Kant's terms, to paraphrase him, we would say that interpretation without perceptions is empty, reception without interpretation is blind. Our knowledge of the external world is based on perception and this means that if there is knowledge in perception then there is also inference in perception. If we were not conscious of the inference, then this would indicate that it is unconscious and happens instantaneously, with the reception by our visual organ of the external things. Therefore, we may say that if we can understand the world by perceiving it, then interpretation is constitutive of our perceptual act. Generally speaking, when we see, we see *as*.

3. As a result, although at the neurophysiological level the same processes occur and the retinal image is the same under standard conditions of perception and normality of the optical apparatus, the epistemic subject interprets the representation from the moment of its production, and this becomes an interpretation that depends on personal expectations, memory, cultural background and other personal characteristics. Therefore, we are entitled to speak of personal knowledge starting from the lowest level of knowledge by acquaintance.

Gombrich takes into account in his research the visual perception of shape, size, distance, and colour. It is obvious that it can be said without epistemic doubts that when we see an object we can see it large or small, close up or far away, in motion or in different lighting conditions. If we recall the distinction between primary and secondary qualities, then there would be reason to distinguish between the form of a thing and its colour, and, respectively, between the perception of form and the perception of colour.

The question arises whether seeing a colour spot of a certain shape would not be a case of seeing *as*, but of seeing. Or, if it is still a case of seeing *as*, is it possible that this time no interpretation is assumed? Isn't this the case where we are acquainted with something? According to the myth of the innocent eye we will say that first we just see a coloured patch of a certain shape and then interpret it and conclude that we see it as such and such. On the contrary, if we claim that interpretation is inherent in the perceptual act, then we will say that we can never have a pure visual experience.

Gombrich mentions several such exemplary cases in which even the apparently purest perceptual experiences involve interpretation. One of these involves the visual experience of colour change as an effect of changing the interpretive framework. If the color changes when we become aware of a particular perspective, this suggests that the earlier experience – mistakenly taken to be "pure" because we were unaware of its interpretive basis – was already shaped by a perspective. In other words, we are always situated within some interpretive framework[5]. Another case is that when we talk about the colour of a shadow (See Gombrich, 1982, p. 30). We realize that we consciously distinguish between shadows and coloured bodies and that, no matter what we are acquainted with, we cannot see shadows as spots of colours since we look at them as shadows. We would immediately become aware of the fact that there is always a prior interpretation guiding the perceptual act if a strange event occurred so that, in full sun, we saw objects without shadows[6].

[5]Gombrich describes this event: "I was looking out of the window, watching for the street car, and I saw through the shrubs by the fence the brilliant red slats of the familiar truck; just patches of red, brilliant scarlet. As I looked, it occurred to me that what I was really seeing were dead leaves on a tree; instantly, the scarlet changed to a dull chocolate brown. I could actually 'see' the change, as one sees changes in a theatre with a shift of lighting. The scarlet seemed positively to fall off the leaves, and to leave behind it the dead brown" (Gombrich, 1977, p. 189).

[6]There is a well-known sequence from Alain Resnais's movie *Last Year at Marienbad*, in which the characters walk through a Cartesian garden, geometrically arranged, in which the shrubs have no shadows. The viewer immediately notices the differences from standard perception.

As a result, we may claim that any perceptual act "depends upon and varies with experience, practice, interests and attitudes" (Goodman, 1976, p. 10). The rejection of the "innocent eye" as an epistemological myth implies acceptance of the idea that the ways we see the world and we make it are shaped by and are relative to our knowledge background, our culture, our expectations, our language, and so on. Therefore, we may assert that "pictures are part of unpictured reality, so the way we see them as pictures is also shaped by our education and training" (Hursthouse, 1992, p. 267).

3. Representing reality: from resemblance to convention

3.1. The very idea of "represent as"

But we do not stop only at this refutation, in the sense proposed by Popper for this kind of epistemic act, of the innocent eye. I think that the main change in the understanding of the pictorial representation of reality consists not only in the revelation and rejection of the myth of the innocent eye but also in the revision of the status of representation so as to understand it not only on the basis of resemblance relations, but also by taking into account conventional elements. Gombrich and Goodman also identify this epistemological shift in the art of painting. In the terms proposed by them, the question is what enables us to recognize a representational picture of x as a representation of x. Following Goodman, Hursthouse distinguishes between two possible answers:

(i) There is a resemblance between x and the picture of x.

(ii) There is a convention to recognize the representational picture of x as a representation of x (Hursthouse, 1992, p. 267).

The difference between the mimetic theory of representation and the conventionalist theory of representation is given by the so-called medium term between the representation bearer and the representational object or between the artwork and the depicted object (See Wang, 2021, p. 85). If the hard core of the mimetic theory is the relation of resemblance, the conventionalist theory of representation holds that any representation is an interpretation based on conventions and that the main problem rests in elucidating the nature of these conventions. The painted image and the object aren't in a direct, unconditional and immediate representational relation, but in an indirect, conditional and mediated one.

When we see a representational picture of Napoleon as a picture of Napoleon we rely on our cultural and knowledge background. According to (i) we may say that we recognize the picture without any intermediary convention because, at a basic level, we see innocently just a painted man that resembles a picture of Napoleon. According to (ii) we will say that, on the contrary, the fact that we are able to recognize a picture of Napoleon is inevitably based on some iconographical conventions. The code of

conventions constrains our interpretation of what we see when we look at a picture and recognize it as a picture of x.[7]

The extreme view that representation depends on conventions is explicitly claimed by Nelson Goodman. He asserts that a realistic representation depends not upon imitation or illusion but upon inculcation, namely, the process of impressing an idea into someone's mind:

> Almost any picture may represent almost anything; that is, given picture and object there is usually a system of representation, a plan of correlation, under which the picture represents the object. How correct the picture is under that system depends upon how accurate is the information about the object that is obtained by reading the picture according to that system. But how literal or realistic the picture is depends upon how standard the system is ... realism is a matter of habit (Goodman, 1976, p. 38).

However, it should be noted (Stoenescu, 2024) that this epistemological relativism accepted by Goodman and subsequently developed by him is absent from Gombrich's view, which is devoted to a Popperian approach to knowledge, focused on the epistemic subject and not on the community of inquiry. Among other things, as it was already noted (Mitrović, 2010), Gombrich excludes the question of a social nature of conventions and, despite some evidence, is not at all tempted by a social constructivist approach.

Gombrich's view agrees in some points with Goodman's extreme conventionalism. However, it should be noted that this epistemological relativism, accepted by Goodman and subsequently developed, is absent from Gombrich, who is devoted to a Popperian approach to knowledge, focused on the epistemic subject and not on the research community. Among other things, Gombrich rejects the question of the social nature of conventions and, despite some evidence, is not at all tempted by a constructivist approach. Gombrich uses the term "the language of art" to express the idea that artists use "systems of schemata", "codes" or "cryptograms", and that we have to know how to decode them so as to see a picture as a picture of x. We may do this if we have previously learned the standard conventions of different types of representations, for example, if we know what a portrait is and have some expectations regarding how it should look.

A famous story about Picasso is instructive about the pitfalls of representation without the use of conventions needed to understand how and why a painted representation is made in a certain way. The story goes that a husband who commissioned Picasso to paint a portrait of his wife was asked by the artist about the almost finished painting, and he was puzzled, stating that the image shown did not resemble his wife. To show how she really looked, he presented Picasso with a photograph. The painter is said to have looked at the photo and said, "Oh, how small it is!".

[7] Hursthouse expresses this idea by analogy with the way we understand language: "To recognize a picture as a picture of x, say a cat on a mat, is like recognizing the words of a sentence as saying that, for example, the cat is on the mat" (Hursthouse, 1992, p. 268).

Therefore, if we accept the distinction between representations of reality based on resemblance and representations based on conventions, then the question is how arbitrary the conventions can be in order to claim that the respective representations stand for something real. I believe that the transition in the history of art from impressionism to expressionism, then to surrealism, can be understood in terms of such a change in our conventions and as a result of their acceptance as part of reality and not just as external elements. I would suggest that this whole transformation in the visual arts is somewhat similar to the debate in the philosophy of science regarding the acceptance of non-Euclidean geometries and their impact on the physical description of the real world. This point marks the difference not only in art but also in science: Henry Poincaré considered the non-Euclidean geometries to be simple conventions, while Albert Einstein took into account the idea that they can be considered representations of reality (See Galison, 2003).

3.2. The case of perspective

Goodman and Gombrich agree on the existence of a conventional element in artistic representation, but they differ when it comes to the status of perspective as a representation meant to more adequately render reality. Goodman himself grasps the difference, quotes Gombrich's idea and develops it further. While Gombrich rejects "the idea that perspective is merely a convention and does not represent the world as it looks", Goodman insists that "pictures in perspective, like any others, have to be read; and the ability to read has to be acquired" (Goodman, 1976, p. 10).

A first problem would be whether the perspective can be linked to the transition from the myth of the innocent eye to a theoretically loaded eye. Then, if we put the problem in this way, the appropriate question would be whether perspective can be considered a free invention or a discovery regarding the ways in which the world can be represented, in both cases not questioning its conventional character. From an epistemic point of view, when viewed from the context of the genesis of these ideas, the situation would be similar to the already mentioned emergence of non-Euclidean geometries, but in a somewhat twisted way. Poincaré proposed a conventionalist perspective on non-Euclidean geometries, while Einstein wondered about the consequences for physical geometry. If Renaissance artists invented or discovered a new convention that made the pictorial representation of the world closer to how it looks in reality, non-Euclidean geometers moved away from the accepted geometric representation and thus challenged theories of the physical world. In both cases, we are dealing with different representations of the world. If in the first case the reality of the world's appearance was recovered with the help of a technique of illusion–the new conventions of perspective as representation–in the second case, a new mapping of the world was proposed by including time as the fourth dimension, therefore a departure from the way in which things are perceived in terrestrial space.

The battle in both cases, in the arts and in these sciences, is for a truthful representation of space in order to better understand reality and its processes. If we take into account the development of Western art "we seem to see the artists initially struggling with perspective, then getting better and better, until they *get it right*."

(Hursthouse, 1992, p. 271). But even in this case, when the paintings seem more and more truthful in relation to perceived reality, it is still about the establishment of cultural conventions. Things and spatial relations will be perceived differently in another cultural space when they are represented pictorially. Gombrich tells the story of a twentieth-century Japanese painter who talked with his father about a square box represented in the Western perspective, and the latter said "What? This box is surely not square, it seems to me very much crooked" (Gombrich, 1977, p. 227).

Therefore, we may say that the Renaissance artists worked within a specific framework or tradition, that of geometrical perspective, and that they tried to represent nature using geometrical conventions in order to present a geometrical figure as it appears. This was, so to speak, a geometrical tradition. Things, their size and shape, as well as the spatial relations between them, are represented according to the rules of Euclidean geometry.

A complementary source of fixing the conventions underlying perspectival representation was geometric optics. It was a period of major interest in the development of optical instruments such as glasses and telescopes, but also in the scientific research of the propagation of light in the form of light rays, understood as straight lines, according to the principles of geometry, hence the new scientific results regarding reflection, refraction and image formation. The artists, beginning with Giotto, learned by trial and error how to use perspective in their paintings and developed new pictorial conventions. The historical context is similar to the late 19th century when Impressionism debuted with Claude Monet's Impression Sunrise, and technical inventions such as photography and cinema, generated new horizons of expectation and new perspectives on the representation of reality. But, surely, this does not mean that science is also entirely conventional.

4. A degree of plasticity

The way out of this crossroads between resemblance and convention can be an interrogative one: "Do we see pictures of *x*s *as* pictures of *x*s because they resemble *x*s or because we have mastered certain conventions? The answer seems to be: well, a bit (or rather, a lot) of both" (Hursthouse, 1992, p. 274).

On the one hand, there is a relation between resemblance and the code of conventions because we may recognise something based on a code. But, at the same time, resemblance isn't entirely relative to systems, codes, or conventions of representations. It is just partially relative to them. Gombrich (1982, pp. 27-28) talks about a degree of plasticity of the ways we see the world, a plasticity which enables us to be taught to see resemblances between pictures and the world with the help of conventions. He notes that when Impressionist paintings were made, many people didn't see them *as* resembling the world. But in the meantime, we have become familiar with Impressionist code, and we begin to see the world as something full of light and we think that the Impressionist paintings accurately represent an aspect of our common world. Expressionism and Surrealism have continued this process: the subject became increasingly autonomous until it came to impose its own conventions on reality and to give the eye

new epistemic functions outside of the purely visual one. With Surrealism, the eye also acquires an introspective, even oneiric and occult function, so that the representational eye is ultimately replaced and reconfigured, as Victor Brauner suggests, into a triple functionality: the "suppressed eye", the "autonomous eye" and the "sublimated eye" (Stoenescu, 2025).

The so-called "degree of plasticity" gives us the capacity to change the way we see the world and to create reality. In "Visual discovery through art" Gombrich (1982) develops the theme of the artist as a creator, already discussed in *Art and Illusion*, especially in the chapter "Pygmalion's Power". Gombrich returns to Aristotle and quotes from his *Poetics*. Aristotle considers that people enjoy looking at accurate representations because this is an instance of our enjoyment of learning. He claims that people enjoy seeing likenesses because they acquire information, for example, they discover that this is a picture of so and so, but Gombrich reverses this point, and suggests that the thrill of learning is found not in recognizing the world in pictures, but in recognizing pictorial effects in the world. Artists propose and create their own vision of the world in their paintings, and we discover and recognise new aspects of the world that were not previously recognised and of which we were not aware or familiar.

Art transforms reality because the representations of reality transfigure it, represent it, or present new aspects, or, moreover, create new ways of representing the world, or of seeing it. Even the most realistic artists re-create the world. Therefore, all acts of seeing are acts of *seeing as*. Ultimately, even the most elementary perceptual act involves isolating and selecting, and to do so means to impose our own vision on the world. The world is seen as categorized into things, but we do not see things only as things. We always see them from a perspective, and our representations become a part of reality. If John Constable dedicated more than one hundred paintings to capturing the sky in different landscapes so as to represent realistically the changing phenomenal appearances of the clouds, the surrealist eye, as Luis Buñuel and Salvador Dali proposed in their film, *Un chien andalou*, refuses to look with a naked eye to the clouds passing in the sky and instead plunges introspectively into a world where we have something else to see. The innocent eye is definitively suppressed.

References

Aristotle, *The Poetics*. (S. H. Butcher, Trans.) Project Gutenberg Ebook, https://www.amherst.edu/system/files/media/1812/The%252520Poetics%252520of%252520Aristotle%25252C%252520by%252520Aristotle.pdf, Retrieved November 2, 2025.

Galison, P. (2003). *Einstein's Clocks, Poincaré's Maps: Empires of Time*. New York: W.W. Norton.

Gombrich, E. H. (1977. *Art and Illusion*. London: Phaidon.

Gombrich, E. H. (1982). *The Image and the Eye: Further Studies in the Psychology of Pictorial Representation*. London: Phaidon.

Goodman, N. (1976). *Languages of Art: An Approach to a Theory of Symbols. Oxford*: Oxford University Press.

Hanson, N. R. (1958). *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge: Cambridge University Press.

Hume, D. (1982). *A Treatise of Human Nature*. (D. G. C. Macnabb, Ed.). Glasgow: Fontana, Collins.

Hursthouse, R. (1992). Truth and representation. In O. Hanfling (Ed.). *Philosophical Aesthetics. An Introduction* (pp. 239–296). Oxford: Blackwell.

Locke, J. *An Essay Concerning Human Understanding*, Pennsylvania State University, Electronic Classics Series. https://www.philotextes.info/spip/IMG/pdf/essay_concerning_human_understanding.pdf, Retrieved October 27, 2025.

Mitrović, B. (2010). A defence of light: Ernst Gombrich, the Innocent Eye and seeing in perspective. *Journal of Art Historiography* (3), 1–30.

Plato (2007), *The Republic*. (D. Lee, Trans.). London: Penguin Books. (References are to the Stephanus numbering).

Pliny the Elder, *Natural History*, Vol. IX, Book XXXV. (H. Rackham, Trans.). Loeb Classical Library, https://www.loebclassics.com/view/pliny_elder-natural_history/1938/pb_LCL394.3.xml?readMode=recto, Retrieved November 10, 2025.

Polanyi, M. (1958). *Personal Knowledge: Towards a Post-Critical Philosophy*. Chicago: University of Chicago Press.

Popper, K. R. (2002). *The Logic of Scientific Discovery*, London and New York: Routledge.

Richmond, S. S. (1994). *Aesthetic Criteria: Gombrich and the Philosophies of Popper and Polanyi*. Amsterdam, Atlanta: Rodopi.

Ryle, G. (1962). The Theory of Meaning. In M. Black (Ed.). *The Importance of Language* (pp. 147–160). Englewood Cliffs, N.J.: Prentice Hall.

Sellars, W. (1956). Empiricism and the Philosophy of Mind. In H. Feigl & M. Scriven (Eds.). *Minnesota Studies in the Philosophy of Science*, I, (pp. 253–329). Minneapolis: University of Minnesota Press.

Stoenescu, C. (2024). Ways of Understanding the Phenomenal. The Cases of Pictorial Representation and Exemplification. In A. I. Mărășoiu & M. Dumitru (Eds.). *Understanding and Conscious Experience. Philosophical and Scientific Perspectives* (pp. 158–176). Routledge Studies in the Philosophy of Science, New York: Routledge, Taylor & Francis Group.

Stoenescu, C. (2025). Ochiul suprarealist de la Dali la ...Hitchcock. Note marginale la expoziția Victor Brauner. Între oniric și ocult. In D. Grusea, C. F. Moraru & H. V. Pâtrașcu (Eds.). Arta gândirii, gândirea artei. Vocea filosofiei. Omagiu profesorului Constantin Aslam la 70 de ani (pp. 136–145). București: Editura Eikon.

Wang, B. (2021). Demystification of the Innocent Eye: Nelson Goodman, Ernst H. Gombrich, and the Limitation of Conventionalism. *Journal of Comparative Literature and Aesthetics* 44 (1). 81–88.

Balkan Journal of Philosophy (2026), vol. 18, issue 1, 32 – 49

UNDERSTANDING BEYOND KNOWLEDGE: FICTION, MODELS, AND COGNITIVE GAIN

Kênio Estrela*

Abstract: This paper examines the epistemic status of understanding and argues that it cannot be adequately reduced to propositional knowledge or to justified true belief. The analysis challenges strictly truth-centered accounts of cognition and defends a conception of understanding grounded in structured interpretive engagement with representational systems. Drawing on recent debates in epistemology and philosophy of science that distinguish understanding from knowledge (e.g., Kvanvig 2003; Grimm 2006; de Regt 2017), the paper situates understanding as a distinct epistemic achievement with its own normative profile, irreducible to belief or information possession. It is argued that scientific models, literary fictions, and contemporary AI-mediated representations systematically rely on idealized or non-literal structures that nonetheless yield genuine epistemic gains. The central claim is that understanding arises through norm-governed practices of interpretation and use, rather than through the mere accumulation of accurate propositions. Scientific models contribute to understanding by isolating explanatory dependencies, fictional narratives afford experiential, perspectival, and meaning-structuring insight, and AI systems support human understanding by facilitating linguistically mediated exploratory and interpretive activities. In each case, departures from literal truth function as epistemically productive constraints rather than as defects. By articulating a unified framework that accounts for these practices, the paper clarifies the relationship between understanding and knowledge and defends the legitimacy of fictional and idealized representations as indispensable components of human sense-making and cognition.

Keywords: understanding, knowledge, scientific models, fiction, representation, epistemology.

1. Introduction

A widespread assumption in epistemology is that genuine cognitive achievement primarily consists in the acquisition of true propositions about the world. On this view, epistemic success is assessed mainly in terms of truth, justification, and reliability, and knowledge is treated as the central epistemic notion. Within such a framework, science appears as the paradigmatic domain of cognition, while art and fiction are commonly regarded as epistemically secondary, associated more with imagination or aesthetic value than with genuine understanding. This assumption underlies many traditional epistemological frameworks, in which understanding is treated either as a derivative notion or as a merely psychological accompaniment to knowledge.

*Escuela de Filosofía, Universidad Finis Terrae, Santiago; kestrela@uft.cl

This orientation reflects what has often been described as a truth-centered or belief-centered conception of epistemic success, dominant in much of twentieth-century epistemology (cf. Kvanvig, 2003; Greco, 2014).

This picture becomes increasingly strained once attention is paid to the actual representational practices employed across different cognitive domains. Scientific inquiry relies extensively on models, idealizations, and abstractions that are known not to be literally accurate descriptions of reality. These representations are nevertheless indispensable for explanation, prediction, and reasoning about complex systems. Their epistemic value does not lie simply in stating true propositions, but in enabling agents to grasp patterns, dependencies, and relations that would otherwise remain opaque through forms of representation that invite interpretation rather than mere belief formation. As has been widely noted in recent philosophy of science, the intelligibility and usability of models often matter more for understanding than their strict descriptive accuracy. Accounts of scientific understanding have therefore increasingly emphasized intelligibility, use, and explanatory traction over representational fidelity (de Regt, 2017; Elgin, 2017).

Parallel considerations arise in the case of art and fiction. Narrative and fictional representations are not ordinarily engaged with as sources of factual information, yet they are widely taken to afford insight into human experience, social relations, and forms of agency. What such representations provide is typically not new knowledge about how the world actually is, but an enhanced capacity to make sense of experiences, perspectives, and possibilities. Their cognitive contribution consists less in the transmission of information than in the reorganization of meaning and attention, a kind of epistemic gain that resists straightforward reduction to propositional content. This form of understanding is often holistic, context-sensitive, and resistant to paraphrasing, a feature frequently highlighted in discussions of understanding as distinct from knowledge (Grimm, 2006; Hills, 2016).

Despite these similarities, philosophical discussions of understanding in science and in art have often developed independently. Accounts of scientific understanding tend to focus on explanation, modeling, and idealization, while discussions of fiction emphasize imagination, narrative structure, and experiential insight. As a result, the deeper commonalities between these practices, particularly their reliance on shared representational conventions, interpretive norms, and criteria of appropriate use, have received comparatively little systematic attention. In particular, the role of norm-governed engagement with representations as a common source of cognitive gain across domains remains underexplored. While recent work has begun to address understanding as a practice-sensitive and gradable epistemic achievement (Baumberger et al., 2017), a unified account spanning science, fiction, and technologically mediated representation is still lacking.

The aim of this paper is to articulate a unified account of understanding that applies across these domains. The central thesis is that many significant forms of cognitive gain are better characterized as instances of understanding rather than as instances of knowledge. Understanding, on the account defended here, is not primarily a matter of possessing accurate beliefs, but of being able to competently interpret,

navigate, and reason within representational frameworks that structure inquiry, interpretation, and reasoning. These frameworks guide attention, constrain inference, and enable agents to explore possibilities, including counterfactual and non-actual scenarios, in ways that are governed by normative expectations rather than by truth alone. This conception aligns with practice-oriented and non-factivist approaches to understanding, according to which epistemic success is grounded in use, intelligibility, and inferential competence rather than belief possession alone (Dellsén, 2017; de Regt, 2017).

From this perspective, fictional and idealized elements are not epistemic shortcomings to be eliminated, but constitutive features of many understanding-generating practices. In science, models abstract and distort in order to render complex systems intelligible. In art and fiction, narrative representations organize experience in ways that make salient patterns of meaning, motivation, and agency. In both cases, understanding depends on the successful uptake of representations within practices that specify how they are to be interpreted, evaluated, and used. Accordingly, epistemic evaluation shifts from assessing representations in isolation to assessing the practices and norms that govern their employment. This shift resonates with accounts that defend the epistemic legitimacy of idealization and fiction when embedded in appropriate normative frameworks (Cartwright, 1983; Elgin, 2017).

The paper proceeds as follows. Section 2 clarifies the conceptual distinction between knowledge and understanding. Section 3 examines the epistemic role of fictional and idealized models in scientific practice. Section 4 turns to art and fiction, analyzing their contribution to cognitive gain. Section 5 develops a general account of understanding as structured and norm-governed cognitive engagement, and Section 6 illustrates this framework through cases drawn from science, fiction, and AI-mediated interpretation. The conclusion reflects on the implications of this account for contemporary epistemology and the theory of representation.

2. Understanding and knowledge: distinct epistemic achievements

Philosophical discussions of cognition have traditionally treated knowledge as the central epistemic achievement. Within this framework, to know is typically understood as possessing true beliefs supported by adequate justification, and epistemic evaluation is largely oriented toward questions of truth, evidence, and reliability. The dominant questions concern how beliefs are formed, whether they are true, and whether they are properly justified or reliably produced. Understanding, when acknowledged at all, has often been treated as a derivative notion, a pragmatic or pedagogical supplement to knowledge rather than a theoretically autonomous epistemic category. This orientation is characteristic of mainstream epistemology, where understanding is frequently subsumed under knowledge or treated as epistemically secondary, lacking independent criteria of success (Kvanvig, 2003; Grimm, 2006).

This traditional orientation reflects an implicit assumption: that the primary aim of cognition is the accurate representation of the world, and that epistemic success consists in matching beliefs to facts. From this perspective, understanding is frequently

conceived as a mere byproduct of knowledge acquisition, a psychological state accompanying the possession of many true beliefs, or a heuristic label for especially well-organized knowledge. Such views leave little conceptual space for understanding as an independent epistemic achievement with its own normative profile. As a result, the epistemic significance of practices that do not primarily aim at literal truth has often been underestimated or misunderstood.

Over the last decades, however, this hierarchy has been increasingly questioned. A growing body of philosophical work argues that understanding constitutes a distinct epistemic achievement, one that cannot be fully reduced to the possession of true propositions. While knowledge is paradigmatically factive and belief-centered, understanding appears to involve grasping relations, patterns, and dependencies that organize information into an intelligible whole (Grimm, 2010; de Regt, 2017). Crucially, such grasp may persist even in the absence of complete, precise, or fully accurate information, so long as the relevant relations can be appropriately interpreted and navigated. This shift has been especially prominent in philosophy of science, where attention to scientific practice has revealed forms of epistemic success not well captured by traditional knowledge-centered models (Dellsén, 2017).

One of the central motivations for this shift comes from reflection on ordinary epistemic practices. It is widely acknowledged that an agent may possess a large body of correct information about a domain and yet fail to understand it. Someone might know many isolated facts about a scientific theory, a historical period, or a social phenomenon while lacking any sense of how those facts fit together, why they matter, or what explains them. Conversely, agents are often credited with understanding even when their representations are simplified, schematic, or known to be inaccurate in certain respects. These everyday judgments suggest that understanding involves more than the accumulation of truths; it involves the ability to make sense of information within a coherent representational perspective. This asymmetry between knowledge and understanding challenges the assumption that epistemic value is exhausted by truth possession (Grimm, 2006).

A prominent line of argument emphasizes that understanding requires insight into how and why certain facts hold, not merely knowledge that they do. Understanding is associated with the ability to explain, to answer “why” and “how” questions, and to see how different elements of a domain depend on one another. This explanatory dimension distinguishes understanding from knowledge conceived narrowly as justified true belief. Knowing that a phenomenon occurs does not by itself amount to understanding why it occurs or how it relates to other phenomena (Lipton, 2004; Grimm, 2010).

This point becomes especially salient when considering the role of integration. Understanding is characteristically holistic: it involves seeing how pieces of information form a coherent structure. An agent who understands a domain can situate particular facts within a broader framework, recognize which factors are explanatorily central, and distinguish relevant from irrelevant details. Knowledge, by contrast, may remain fragmented. One can know many truths without possessing the unifying perspective that understanding provides, or without being able to deploy those truths

within a meaningful inferential pattern. This contrast supports the idea that understanding has an essentially structural dimension that is absent from many accounts of knowledge (Kvanvig, 2003).

These considerations have led many philosophers to treat understanding as a higher-order epistemic achievement. Rather than consisting in the possession of individual beliefs, understanding involves the organization, coordination, and application of information within a structured representational framework. This framework enables agents to reason effectively, to make sense of new information, and to extend their grasp of the domain beyond what is explicitly known, including by guiding interpretation in novel or non-standard cases (Baumberger, Beisbart, and Brun, 2017).

The distinction between knowledge and understanding has been particularly influential in philosophy of science. Scientific practice provides numerous examples in which understanding is achieved through representations that are known to be idealized or even false. Scientific models often deliberately simplify complex systems, omit causal factors, or introduce fictional elements in order to make patterns and dependencies salient. These representations are not typically regarded as epistemic failures. On the contrary, they are central to explanation, prediction, and theoretical reasoning. Their epistemic success is commonly attributed to their capacity to render phenomena intelligible rather than to their strict descriptive accuracy (Cartwright, 1983; de Regt and Gijsbers, 2017).

Scientific understanding, in this context, is commonly associated with intelligibility. A theory or model is said to promote understanding when it renders a phenomenon intelligible to its users, enabling them to reason about it, to answer explanatory questions, and to explore counterfactual scenarios. Importantly, intelligibility is not a purely semantic property of representations, nor does it require strict truth at every point. It depends on how representations function within a practice and on the skills agents bring to their use, including their capacity to interpret representational elements in accordance with shared norms (de Regt, 2015; 2017).

From this perspective, understanding involves epistemic abilities rather than merely epistemic states. To understand a phenomenon is to be able to use a theory or model competently: to manipulate it, to draw inferences from it, to recognize its scope and limitations, and to apply it to novel situations. These abilities may be exercised even when the underlying representation includes idealizations or known inaccuracies. What matters is not perfect correspondence with reality, but the capacity of the representational system to support reliable and illuminating reasoning through appropriate modes of use (Dellsén, 2017).

Closely related to this emphasis on intelligibility is the idea that understanding has a counterfactual dimension. An agent who understands a phenomenon is typically able to answer questions about how it would behave under different conditions, what would happen if certain factors were altered, or which aspects of the system are responsible for particular outcomes. Such counterfactual competence goes beyond knowing what actually happens; it involves grasping the structure of dependencies that govern the phenomenon as they are represented within a model or framework, a point

emphasized in accounts that link understanding to the ability to explore counterfactual dependencies through models (Kuorikoski, 2011).

This conception of understanding helps explain why it is compatible with epistemic imperfection. Models that contain false assumptions can still support understanding if they accurately capture relevant dependencies and allow users to reason effectively about the system. In this sense, understanding tolerates a degree of representational inaccuracy that knowledge, strictly construed, does not. This tolerance is not a weakness, but a reflection of the different epistemic roles played by understanding and knowledge (Elgin, 2017).

The distinction between knowledge and understanding also has normative implications. Knowledge is typically evaluated in terms of truth and justification: a belief either meets the relevant standards or it does not. Understanding, by contrast, admits of degrees and is sensitive to context. One may understand a phenomenon better or worse, depending on the richness of the representational framework, the range of counterfactuals one can handle, and the purposes for which the understanding is deployed. This gradability further distinguishes understanding from knowledge (Baumberger, 2019).

Seen in this light, understanding occupies a distinct epistemic space between mere belief and full theoretical knowledge. It is structured, relational, and practice-oriented. It depends on mastery of representational frameworks and inferential norms rather than on the passive possession of truths. Understanding is achieved through engagement: through using models, narratives, diagrams, or other representational devices to make sense of complex domains in ways that are governed by norms of interpretation rather than by truth alone.

This reconceptualization has important consequences for how fictional and idealized representations are evaluated. If understanding does not require strict factual accuracy at every representational level, then the presence of fictional elements in epistemic practices need not be regarded as epistemically suspect. On the contrary, such elements may be essential for highlighting salient relations, constraining interpretation, and supporting exploratory reasoning. Fiction and idealization can function as cognitive tools rather than as epistemic obstacles.

Moreover, this account helps explain why understanding is often robust under revision. Scientific theories change, models are replaced, and representations are refined, yet the understanding they afford may persist or even deepen. This persistence suggests that understanding is not tied to any particular set of propositions, but to the broader representational practices through which agents engage with a domain.

The distinction between knowledge and understanding thus sets the stage for the analyses that follow. If understanding is a distinct epistemic achievement grounded in structured engagement with representations, then it becomes possible to reassess the epistemic roles of scientific models and works of fiction. The next section examines scientific modeling as a paradigmatic case in which fictional and idealized representations contribute centrally to understanding.

3. Scientific models and epistemic fiction

Scientific practice relies extensively on models that simplify, idealize, or deliberately distort aspects of reality. From frictionless planes in classical mechanics to ideal gases in thermodynamics and highly simplified agents in economics, many of the central representational tools of science are known to be literally false. They do not aim to describe the world in full detail, nor do they aspire to capture every causal factor at work in the phenomena they represent. Yet these models are not treated as epistemic failures. On the contrary, they are widely regarded as indispensable for explanation, prediction, and, most importantly, scientific understanding (Cartwright, 1983; de Regt, 2017).

This situation poses an apparent puzzle. If epistemic success were exhausted by the accurate representation of facts, the pervasive reliance on idealized and inaccurate models would be difficult to justify. Scientific inquiry would seem systematically compromised by its own methods. The persistence and centrality of modeling practices therefore motivate a reassessment of the epistemic standards by which scientific representations are evaluated, shifting attention from accuracy alone to the roles representations play within inquiry (Mizrahi, 2012; Dellsén, 2017).

Rather than viewing models as incomplete descriptions awaiting correction, contemporary philosophy of science increasingly emphasizes their functional role within scientific inquiry. Models are not primarily mirrors of reality, but representational instruments designed to make complex systems intelligible. They enable scientists to isolate dependencies, explore counterfactual scenarios, and identify explanatory patterns that would remain opaque if all details were retained. Their epistemic value lies less in their literal truth than in what they enable users to do with them, namely, to reason, explain, and explore within a structured representational space, including the exploration of counterfactual dependencies and the generation of intelligibility despite idealization (Kuorikoski, 2011; de Regt and Gijsbers, 2017).

On this view, scientific models often function as disciplined or constrained fictions. They introduce simplified systems, ideal conditions, or counterfactual assumptions that are known not to obtain in the world. These fictional elements are not arbitrary inventions. They are governed by theoretical commitments, empirical constraints, and methodological norms that regulate how the model is constructed, interpreted, and applied. The departure from literal accuracy is deliberate and controlled, serving specific cognitive purposes by guiding interpretation rather than by asserting facts (Elgin, 2017).

The epistemic contribution of such models is best understood in terms of intelligibility. A model promotes understanding when it renders a phenomenon cognitively tractable for its users. This involves more than the communication of isolated truths. It requires that the model support coherent reasoning, explanation, and exploration. A model is intelligible insofar as scientists can use it to answer "how" and "why" questions, to anticipate how the system would behave under hypothetical changes, and to situate particular observations within a broader explanatory framework that organizes significance and relevance (de Regt, 2015; de Regt, 2017).

Importantly, intelligibility is not a purely representational property. It depends on the interaction between the model and the skills of its users. To understand a phenomenon through a model is to be able to manipulate it, to recognize the significance of its assumptions, and to draw appropriate inferences from it. These abilities may be exercised even when the model includes assumptions that are strictly false. What matters epistemically is not perfect correspondence with reality, but the model's capacity to guide competent reasoning within accepted norms of interpretation and use (Grimm, 2010; Dellsén, 2017).

This perspective helps explain why idealization is often epistemically productive. By omitting or distorting certain features of a system, a model can make salient relations that would otherwise be obscured. Too much detail can hinder understanding by masking patterns and dependencies. Simplification, in this context, is not a loss but a gain, insofar as it enables the isolation of explanatorily relevant dependencies and the stabilization of interpretive use across contexts (Cartwright, 1983; Kuorikoski, 2011).

From this standpoint, distortion is not an epistemic defect but a methodological resource. A model may misrepresent the exact behavior of a system while still accurately capturing how key variables interact. It may fail to describe what actually happens in every circumstance while nonetheless illuminating why certain outcomes occur or how they would change under different conditions. The understanding afforded by the model resides in these structural insights, not in the literal truth of each assumption taken in isolation (Elgin, 2017).

This account also sheds light on the resilience of scientific understanding under theoretical revision. As empirical knowledge advances, models are refined, replaced, or abandoned. Yet the understanding they provide often persists or even deepens. This persistence suggests that the cognitive gain associated with a model is not exhausted by the particular propositions it encodes. Instead, it is grounded in the inferential practices, explanatory perspectives, and counterfactual reasoning the model makes possible across successive representational frameworks (Dellsén, 2016).

The role of fiction in scientific modeling further highlights the importance of conventions and shared interpretive norms. Scientists must coordinate on how a model is to be used, which idealizations are to be taken seriously, and which aspects of the representation are to be bracketed. A frictionless plane, for example, is not taken to deny the existence of friction, but to suspend it for the sake of isolating other dependencies. Understanding arises not from the model alone, but from a socially regulated practice that governs how the fictional scenario is connected to real-world phenomena through established interpretive rules (de Regt and Gijsbers, 2017).

This underscores the normative and practice-dependent character of scientific understanding. Understanding is not a private mental state reducible to belief possession. It is an achievement embedded in communal practices of modeling, interpretation, and explanation. The epistemic success of a model depends on how it is used within these practices, not merely on its representational accuracy or its correspondence to isolated facts (Baumberger, Beisbart, and Brun, 2017).

Seen in this light, scientific models exemplify a broader epistemic pattern. Understanding is achieved through structured engagement with representational systems that are neither fully literal nor purely subjective. Fictional elements are harnessed in a controlled and norm-governed manner to facilitate insight, orientation, and explanatory control. Far from undermining rational inquiry, such elements are integral to the ways in which science makes the world intelligible.

As the next section argues, this epistemic pattern is not unique to science. It also characterizes the cognitive contributions of art and fiction, where understanding likewise emerges from imaginative engagement guided by shared conventions rather than from the transmission of literal truths.

4. Art, fiction, and cognitive Gain

While scientific models provide a paradigmatic case of epistemically productive idealization, similar patterns of cognitive gain can be observed in the domain of art and fiction. Unlike scientific representations, fictional narratives and artworks are not typically evaluated in terms of truth or empirical accuracy. They are openly non-factive and are engaged with under conventions that suspend literal belief. Yet despite this, works of fiction are widely credited with yielding genuine insight into human experience, social structures, moral conflict, and forms of agency (Elgin, 2017; Currie, 2020).

This apparent tension has long posed a challenge for epistemology. If epistemic value is tied essentially to truth, then the cognitive significance commonly attributed to fiction would appear misguided or merely metaphorical. However, a growing body of work in aesthetics and epistemology argues that fiction can contribute to understanding without functioning as a source of propositional knowledge. The relevant epistemic achievement is not the acquisition of new facts, but the reorganization of attention, perspective, and interpretive frameworks through which agents make sense of the world (Gaut, 2003; Elgin, 2017).

Fictional narratives operate by constructing imagined scenarios, characters, and events that are not intended to describe reality as it is. Instead, they present structured representations that invite readers or viewers to explore possibilities, counterfactuals, and patterns of significance. What is epistemically valuable in such engagement is not belief in the content of the fiction, but the disciplined imaginative activity it supports. Through narrative structure, point of view, and symbolic articulation, fiction enables agents to grasp relationships, motivations, and constraints that may be difficult to articulate or apprehend through literal description alone (Currie, 2010; Hills, 2016).

In this respect, fiction shares important functional similarities with scientific models. Both employ departures from literal accuracy in order to enhance intelligibility. Just as idealized models isolate causal dependencies by bracketing irrelevant factors, fictional narratives isolate experiential, moral, or social structures by abstracting away from empirical contingency. The resulting representations are not evaluated primarily

in terms of truth, but in terms of how effectively they support interpretation, reflection, and understanding within established practices of use.

The epistemic contribution of fiction is therefore best understood as perspectival and structural rather than informational. Works of literature, film, or other narrative media can deepen understanding by reorganizing how agents perceive and interpret familiar phenomena. They can make salient patterns of action, forms of vulnerability, or normative tensions that are otherwise overlooked, not by stating them explicitly, but by embedding them within coherent narrative structures. This kind of understanding is holistic and context-sensitive, resisting reduction to a set of discrete propositions (Elgin, 2017; Currie, 2020).

Crucially, the cognitive gains afforded by fiction depend on shared interpretive conventions. Readers do not approach fictional narratives as assertions about the actual world, but as invitations to imagine under specific constraints. These conventions regulate what kinds of inferences are appropriate, which features of the narrative are salient, and how fictional content may legitimately inform real-world reflection. Without such norms, imaginative engagement would collapse into arbitrary fantasy, and fiction would lose its capacity to generate understanding.

This norm-governed character explains why not all fictions are equally epistemically valuable. Just as poorly constructed scientific models fail to support understanding, incoherent or norm-violating narratives fail to yield meaningful insight. Epistemic evaluation in the aesthetic domain therefore concerns the adequacy of representational structure, coherence, and interpretive affordances rather than factual accuracy. Understanding arises when imaginative engagement is guided by stable conventions that support disciplined exploration rather than unconstrained invention (Gaut, 2003; Elgin, 2017).

The role of counterfactual reasoning is particularly salient in fiction. Narrative representations invite agents to consider how individuals might act under different circumstances, how social structures shape agency, or how values conflict in non-actual scenarios. This counterfactual exploration expands the space of intelligibility by allowing agents to grasp dependencies and possibilities without requiring empirical instantiation. As in scientific modeling, the epistemic payoff lies in the ability to navigate these possibilities in a principled way, not in their literal truth.

Understanding through fiction is also compatible with epistemic imperfection. Fictional narratives may exaggerate, simplify, or stylize aspects of human life. Such distortions do not undermine their epistemic value so long as they serve to illuminate structurally relevant relations. Indeed, exaggeration and stylization often play a role analogous to idealization in science, enabling insight by foregrounding what matters most for interpretive purposes (Elgin, 2017).

These considerations support a broader claim about understanding as an epistemic achievement. Understanding does not require that representations be true, but that they be usable within norm-governed practices that support coherent interpretation, counterfactual reasoning, and reflective application. Art and fiction exemplify

this mode of epistemic engagement particularly clearly, revealing how cognitive gain can be achieved through imaginative structures rather than through belief formation.

By situating fiction alongside scientific modeling, this analysis challenges the sharp epistemic divide often drawn between science and art. Both domains employ representational fictions to generate understanding, differing primarily in their aims, constraints, and modes of evaluation rather than in their epistemic legitimacy. Recognizing this continuity prepares the ground for a general account of understanding that treats structured engagement with representations as its core feature.

The next section develops this account explicitly, articulating understanding as a form of structured and norm-governed cognitive engagement that applies across scientific, artistic, and technologically mediated contexts alike.

5. Understanding as structured engagement

The preceding sections suggest a common epistemic pattern underlying both scientific modeling and artistic fiction. In neither case is cognitive success best explained in terms of the accumulation of true propositions. Instead, understanding emerges from a form of engagement with representational systems that guide attention, structure interpretation, and enable meaningful inference. This approach aligns with recent work that treats understanding as a distinct epistemic achievement irreducible to belief or information possession (Grimm, 2014; Elgin, 2017). This section develops a general account of understanding along these lines, characterizing it as a form of structured cognitive engagement rather than as a mental state defined by belief or truth alone.

On this view, understanding is an epistemic achievement that involves the active coordination of multiple cognitive elements: representational artifacts, interpretive conventions, background assumptions, and inferential practices. To understand a phenomenon is not merely to know certain facts about it, but to be able to situate those facts within an organized framework that renders their relations, relevance, and implications intelligible. This emphasis on organization and relational grasp reflects objectual and holistic accounts of understanding, according to which understanding targets structured domains rather than isolated propositions (Kvanvig, 2003; Khalifa, 2017). Such frameworks are typically provided by models, narratives, diagrams, or symbolic systems that impose structure on domains that would otherwise remain cognitively opaque by making salient what counts as relevant, explanatory, or significant.

This conception departs from accounts that locate understanding primarily in an internal psychological state or in the possession of particular beliefs. While mental attitudes undoubtedly play a role, understanding cannot be reduced to what an individual believes in isolation. Rather, it is essentially relational and practice-dependent. It depends on the agent's ability to engage competently with external representational systems and to navigate the norms governing their use. In this respect, understanding is better characterized in terms of epistemic abilities and competencies than in terms

of belief states (Grimm, 2012; de Regt, 2017). Understanding is thus as much a matter of epistemic participation as of mental endorsement, involving mastery of ways of using representations rather than mere assent to their content.

A central feature of structured engagement is its reliance on conventions. Representational systems function epistemically only insofar as users share norms governing how those systems are to be interpreted and employed. In scientific modeling, such conventions determine which elements of a model are idealized, which are intended to track features of the target system, and how inferences drawn within the model may legitimately be projected onto reality. In art and fiction, analogous conventions regulate imaginative uptake, the suspension of literal belief, and the relevance of fictional content to real-world concerns by specifying how meaning is to be taken rather than what is literally asserted. This norm-governed dimension of representation has been emphasized in accounts of scientific understanding centered on intelligibility and use (de Regt and Dieks, 2005; de Regt, 2017).

Importantly, these conventions do not merely constrain interpretation; they actively enable understanding. By stabilizing expectations and coordinating inferential practices, conventions allow fictional or idealized representations to support coherent reasoning, comparison, and reflection. Without such normative structure, imaginative engagement would collapse into arbitrary fantasy, and models would lose their epistemic grip on the world. Understanding thus depends less on representational accuracy at the level of individual elements than on the stability and appropriateness of the interpretive framework within which those elements function (Elgin, 2017). It is within such frameworks that representations acquire epistemic significance.

Another defining feature of understanding as structured engagement is its counterfactual dimension. Both scientific models and fictional narratives invite users to explore possibilities, alternatives, and variations that need not be actual. The capacity to reason counterfactually, to ask how a system would behave under different conditions, or how an agent might respond in a different context, is central to explanatory and interpretive insight. Counterfactual competence is widely regarded as a hallmark of understanding rather than mere knowledge (Woodward, 2003; Khalifa, 2017). Fictional elements facilitate this process by loosening the constraints of actuality while preserving structurally relevant relations.

This account also helps explain why understanding is often resilient in the face of error. A model may incorporate false assumptions, and a fictional narrative may rely on implausible or even impossible scenarios, yet both can still generate genuine understanding if they succeed in organizing experience in a coherent and explanatorily productive way. What matters is not the literal truth of each representational component, but the overall capacity of the system to support stable interpretation, counterfactual reasoning, and guided application. This tolerance for idealization and distortion has been widely recognized in contemporary accounts of scientific understanding (de Regt, 2017; Elgin, 2017).

Understanding, therefore, should not be conceived as a static epistemic state, but as a dynamic and ongoing practice. It involves skills, dispositions, and interpretive

competencies exercised through sustained engagement with representational artifacts. This practical and normative dimension distinguishes understanding from knowledge conceived narrowly as justified true belief. While knowledge can, at least in principle, be possessed passively, understanding requires active participation in a structured cognitive activity governed by shared norms and standards of use that regulate how representations are taken to mean and how they may be extended to new cases.

By framing understanding in this way, the apparent epistemic tension between fiction and cognition dissolves. Fictional and idealized representations are not obstacles to understanding, but integral components of many understanding-generating practices. When embedded within stable conventions and employed for disciplined exploration, they enable forms of insight that would be inaccessible through strictly literal or purely propositional means alone.

The final section of the paper applies this account to a set of illustrative cases drawn from science, fiction, and contemporary representational technologies. These cases clarify how structured engagement operates across different domains and show how a unified conception of understanding can illuminate the epistemic roles of models, narratives, and mediated representations alike.

6. Illustrative cases: science, fiction, and AI-mediated interpretation

The account of understanding developed in the previous sections gains depth and precision when examined through concrete epistemic practices. This section considers three domains in which structured engagement plays a central role in generating understanding: scientific modeling, literary fiction, and AI-mediated interpretation. Although these practices differ in aims, methods, and institutional settings, they share a reliance on fictional, idealized, or non-literal representations governed by stable conventions of use. Recent work in philosophy of science and aesthetics has increasingly emphasized these shared structural features (de Regt, 2017; Elgin, 2017). Examining these cases helps clarify how understanding arises through engagement with representational systems rather than through the mere possession of true beliefs or the passive reception of information.

6.1. Scientific models and idealization

Scientific models routinely employ idealizations and known falsehoods. Frictionless planes, perfectly rational agents, isolated systems, and homogeneous populations are familiar examples across physics, economics, and biology. From the standpoint of a strictly truth-centered epistemology, such representations appear epistemically deficient, since they misdescribe the systems they are meant to represent. Yet in scientific practice, these models are treated as central vehicles of understanding rather than as provisional approximations awaiting elimination.

On the present account, the epistemic role of such models is best explained in terms of structured engagement. Models generate understanding by enabling scientists to explore dependency relations, isolate explanatorily relevant factors, and reason

counterfactually about how systems would behave under varying conditions. This functional conception of models is widely defended in contemporary philosophy of science, particularly in accounts that emphasize intelligibility over literal accuracy (de Regt and Dieks, 2005; Weisberg, 2013). What is gained is not primarily a set of propositions believed to be true, but a practical grasp of how to navigate a space of possibilities in a disciplined and informative manner according to shared inferential standards.

Importantly, this understanding can persist even when the idealizations involved are explicitly acknowledged as false. A physicist does not believe that friction is absent, nor does an economist believe that agents are perfectly rational. What matters epistemically is mastery of the inferential practices associated with the model: knowing when it applies, how its assumptions constrain its use, and how its results can be related back to empirical systems. Idealization thus functions as an epistemic strategy rather than as a representational defect (Elgin, 2017; Frigg and Nguyen, 2020).

This case illustrates a key feature of understanding as structured engagement: epistemic success depends less on representational fidelity at the level of individual assumptions than on the stability and productivity of the interpretive framework within which reasoning takes place. Scientific understanding emerges from participation in a norm-governed practice of model use, not from assent to the literal content of the model itself taken in isolation.

6.2. Literary fiction and experiential understanding

A parallel structure can be observed in the case of literary fiction. Readers typically do not approach novels, plays, or films with the expectation of acquiring factual knowledge about the world. Nevertheless, engagement with fiction is widely regarded as cognitively valuable, yielding understanding of human motivation, social interaction, moral conflict, and emotional life.

This cognitive gain is not reducible to the extraction of general truths from narrative content. Rather, fiction affords structured experiential access to possibilities that may be rare, inaccessible, or ethically unavailable in real life. Philosophers of art and literature have emphasized that such understanding arises through imaginative engagement governed by narrative and genre conventions (Currie, 1995; Gaut, 2003). By imaginatively inhabiting fictional scenarios, readers explore perspectives and situations in a controlled environment shaped by narrative conventions. These conventions guide attention, regulate interpretation, and constrain imaginative engagement in ways that support coherent reflection and meaningful comparison with real-world experience.

From the standpoint of structured engagement, fiction functions as an epistemic laboratory. It allows readers to test responses, evaluate choices, and recognize patterns of agency and vulnerability without committing to the belief that the represented events actually occurred. The understanding gained is holistic and non-propositional, involving sensitivity to salience, context, and relevance rather than the endorsement of explicit generalizations (Elgin, 2017; Currie, 2020).

This perspective helps explain why fictional understanding is often resistant to paraphrasing. What is grasped through narrative engagement cannot always be captured by a list of statements, even when such statements are true. The epistemic value of fiction lies in its capacity to organize experience and cultivate interpretive skill, not in its contribution to the stock of factual knowledge considered independently of context.

6.3. AI systems and interpretive mediation

Recent developments in artificial intelligence offer a contemporary extension of this framework. Large language models and other AI systems do not possess beliefs, intentions, or understanding in any robust philosophical sense. Nonetheless, they increasingly mediate human understanding by generating representations with which users engage interpretively through linguistic interaction.

AI-generated outputs often involve simplifications, generalizations, and occasional errors. Evaluated purely in terms of truth or reliability, such systems may appear epistemically fragile. Yet in practice, they can support understanding when embedded within appropriate conventions of use. Recent philosophical work on scientific representation and cognitive tools emphasizes that epistemic value often lies in how representational artifacts are used rather than in their intrinsic accuracy (de Regt, 2017; Frigg and Nguyen, 2020).

On the present account, the epistemic contribution of AI systems resides not in the representations they produce in isolation, but in the structured interaction they enable. When users engage reflectively with AI-generated content, testing it against background knowledge and situating it within broader interpretive frameworks, cognitive gain can occur despite imperfections in output. Understanding arises from the human practice of use, not from the machine's internal processes or its capacity to store or retrieve information.

This case reinforces a central claim of the paper: understanding is not located solely in representational accuracy or informational content, but in norm-governed practices of engagement. Even representations that are partially fictional, idealized, or unreliable can contribute to understanding when used within stable and reflective interpretive conventions that regulate how they are taken to mean and how they are integrated into ongoing inquiry.

7. Conclusion: understanding as structured cognitive engagement

This paper has argued that understanding constitutes a distinct epistemic achievement that cannot be reduced to the possession of true propositions or justified belief. By examining epistemic practices across science, fiction, and contemporary forms of technological mediation, it has shown that understanding often arises through engagement with representations that are idealized, fictional, or partially inaccurate, yet systematically constrained by norms of use and interpretation.

The framework proposed here characterizes understanding as structured cognitive engagement. On this view, epistemic success does not hinge on strict factual correspondence at the level of individual representations, but on the presence of normative constraints, interpretive conventions, and inferential affordances that guide cognitive interaction. Scientific models, literary narratives, and AI-generated representations exemplify this structure in different ways. Each relies on controlled departures from literal truth to facilitate intelligibility, orientation, and insight within shared practices of sense-making, a point increasingly emphasized in contemporary accounts of understanding and intelligibility (de Regt, 2017; Elgin, 2017).

This account helps explain why fictional and idealized elements can play a legitimate epistemic role without undermining rational inquiry. In science, idealizations enable the isolation of causal patterns and explanatory relations that would otherwise remain obscured. In fiction, narrative constructions afford experiential access to psychological, moral, and social possibilities that resist purely propositional articulation. In AI-mediated contexts, simplified and probabilistic outputs support exploration, comparison, and conceptual clarification when embedded within appropriate interpretive practices. In all cases, understanding emerges from use and engagement rather than from representational content considered in isolation or assessed solely by its truth conditions.

By emphasizing engagement over belief, the framework also clarifies the relationship between understanding and knowledge. Knowledge remains indispensable where accuracy, justification, and reliability are the primary epistemic aims. Understanding, however, operates with a different normative profile. It tolerates non-literal representations, counterfactual reasoning, and controlled distortions, provided that these contribute to coherent cognitive orientation and explanatory competence. In this sense, understanding is compatible with epistemic humility: one can understand a phenomenon without claiming exhaustive or final knowledge of it, or full propositional mastery of its details.

The analysis contributes to ongoing debates in philosophy of science, aesthetics, and epistemology by offering a unified account of epistemic practices that are often treated in isolation. It suggests that the apparent divide between science and art reflects differences in representational aims rather than differences in epistemic value. Both domains employ fictions and conventions to cultivate understanding, and both can be evaluated according to how effectively they structure cognitive engagement and guide interpretation.

Finally, the proposed framework has implications for contemporary epistemology in an age of increasing technological mediation. As AI systems become integral to scientific research, education, and public discourse, assessing their epistemic role requires moving beyond narrowly truth-centric models. Understanding how such systems can enhance, rather than obscure, human cognition depends on recognizing the normative practices that govern their use and the interpretive competencies of their users.

Understanding beyond knowledge, the paper concludes, is not an epistemic deficiency but a fundamental mode of human cognition: one through which agents make sense of the world by engaging with representations that are intelligible, disciplined, and normatively constrained, even when they are not strictly true.

References

Baumberger, C. (2019). Explicating Objectual Understanding: Taking Degrees Seriously. *Journal for General Philosophy of Science*, 50, 367–388.

Baumberger, C., C. Beisbart, and G. Brun. (2017). What Is Understanding? An Overview of Recent Debates in Epistemology and Philosophy of Science. In *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, edited by Stephen R. Grimm, Christoph Baumberger, and Sabine Ammon. New York: Routledge.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Currie, G. (1995). *Image and Mind: Film, Philosophy and Cognitive Science*. Cambridge: Cambridge University Press.

Currie, G. (2010). *Narratives and Narrators: A Philosophy of Stories*. Oxford: Oxford University Press.

Currie, G. (2020). *Imagining and Knowing: The Shape of Fiction*. Oxford: Oxford University Press.

Dellsén, F. (2016). Scientific Progress: Knowledge versus Understanding. *Studies in History and Philosophy of Science*, 56, 72–83.

Dellsén, F. (2017). Understanding without Justification or Belief. *Ratio*, 30(3), 239–254.

de Regt, H. W. (2014). Visualization as a Tool for Understanding. *Perspectives on Science*, 22(3), 377–396.

de Regt, H. W. (2015). Scientific Understanding: Truth or Dare? *Synthese*, 192, 3781–3797.

de Regt, H. W. (2017). *Understanding Scientific Understanding*. Oxford: Oxford University Press.

de Regt, H. W. and D. Dieks. (2005). A Contextual Approach to Scientific Understanding. *Synthese*, 144, 137–170.

de Regt, H. W., and Victor Gijsbers. 2017. "How False Theories Can Yield Genuine Understanding." *In Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, edited by Stephen R. Grimm, Christoph Baumberger, and Sabine Ammon. New York: Routledge.

Elgin, C. Z. (2017). *True Enough*. Cambridge, MA: MIT Press.

Frigg, R. and J. Nguyen. (2020). *Modelling Nature: An Opinionated Introduction to Scientific Representation*. Cham: Springer.

Gaut, B. (2003). Art and Knowledge. In *The Oxford Handbook of Aesthetics*, edited by Jerrold Levinson. Oxford: Oxford University Press.

Grimm, S. R. (2006). Is Understanding a Species of Knowledge? *British Journal for the Philosophy of Science*, 57(3), 515–535.

Grimm, S. R. (2010). The Goal of Explanation. *Studies in History and Philosophy of Science*, 41 (4), 337–344.

Grimm, S. R. (2014). Understanding as Knowledge of Causes. In *Virtue Epistemology Naturalized*, edited by Abrol Fairweather, 329–345. Cham: Springer.

Grimm, S. R. (2017). Understanding and Transparency. In *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, edited by Stephen R. Grimm, Christoph Baumberger, and Sabine Ammon. New York: Routledge.

Hills, A. (2016). *Understanding Why*. Oxford: Oxford University Press.

Khalifa, K. (2017). *Understanding, Explanation, and Scientific Knowledge*. Cambridge: Cambridge University Press.

Kuorikoski, J. (2011). Simulation and the Sense of Understanding. In *Models, Simulations, and Representations*, edited by Paul Humphreys and Cyrille Imbert. New York: Routledge.

Kvanvig, J. L. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge: Cambridge University Press.

Lipton, P. (2004). *Inference to the Best Explanation*. 2nd ed. London: Routledge.

Mizrahi, M. (2012). Idealizations and Scientific Understanding. *Philosophical Studies*, 160(2), 237–252.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Balkan Journal of Philosophy (2026), vol. 18, issue 1, 50 – 65

UNDERSTANDING AND SUBJECTIVITY

André W. Carus*

Abstract: Where is understanding located? However this question is answered, it is widely assumed that ultimately individual minds do the understanding. But cognitive science has cast significant doubt on the idea of an individual mind, so how can this work? This paper uses the responses of Ismael (2007, 2016) and Hofstaedter (2007) to these cognitive-science challenges as a basis for developing a conception of individual understanding. It then also suggests that the same conception can trace both objective understanding (of the natural and social worlds) and subjective understanding (of persons, values, reflections on values in art and literature) to a common source in the individual mind's capability of allocating attention to investment in its mental life. In both cases, understanding is seen to be an extended process, requiring persistent attention, not an isolated event.

Keywords: subjective understanding, scientific understanding, concept of self, Dennett, Daniel, Ismael, Jenann, world models, cognitive background corpus.

Where is understanding actually located? Does it reside in a scientific-disciplinary community? Or only in its constituent individuals? Or both? The shift in attention from knowledge to explanation to understanding, in recent years, has mostly left these questions aside. Different answers to them are assumed, sometimes explicitly, by different writers about understanding, but the question itself, about the locus of understanding, rarely arises.

It is widely agreed that in principle, understanding resides in individual minds. It is generally conceded, as Baumberger et al. (2017, p. 6) put it, that "understanding is a cognitive achievement that can be ascribed to an agent," an individual mind. Schurz and Lambert (1994) also proceeded from this assumption, but their theory of understanding sought independence from psychological considerations; the "cognitive background corpus" of the individual agent (the fitting of a particular phenomenon into which constitutes understanding, in their view) turns out to consist in (an appropriately abridged version of) the entire body of objectively available scientific knowledge and lacks the sort of embeddedness in tacit and informal knowledge a psychologically realistic account would require (cf. Tall, 2013).

Henk de Regt's conception of contextual understanding (de Regt, 2017; 2020; 2025) abandons any such quest for objective and non-perspectival understanding, and though still locating ultimate understanding in the constituent individuals, it regards the scientific communities they compose as the primary bearers of understanding. Scientific understanding "is a community achievement: scientific progress is possible only if

*LMU Munich; awc@awcarus.com

there is a shared understanding within a community" (de Regt, 2017, p. 141). Understanding is located in the minds of individual scientists, but their common socialization within a particular discipline or field leads to considerable overlap in shared understanding: "a human being is always part of a context – a historical and social context, for example. For scientists the context is in important ways determined by their disciplinary education, by their background knowledge and by the state of the art in their field. So, whether or not a scientific theory can provide understanding is partly dependent on who the understanders are, and what their context is" (de Regt, 2025, pp. 27–8). In the terms of Schurz and Lambert, members of a discipline share a common "cognitive background corpus." This applies not just to their common fluency in the mathematical *languages* of the field in question, but also, for de Regt, the overlap in a great deal of the informal and "tacit" knowledge about what the live issues are, what is important, who to pay attention to and who to ignore, how things fit together, and so on (de Regt, 2017, pp. 27-9). This "background corpus" includes as much knowledge-how as knowledge-that – i.e. we can't restrict it to a strictly *cognitive* background corpus, as Schurz and Lambert do.

In de Regt's latest contribution, the locus of understanding has shifted somewhat further toward the individual: "understanding involves a thinking subject, an 'understander.' It is human beings who understand…" (de Regt, 2025, p. 27). Moreover, de Regt offers an account of musical understanding as an illuminating parallel to his account of scientific understanding, and here we are unquestionably in at least partly subjective territory. His account of understanding also departs, therefore, from previous discussions (not just Schurz and Lambert but also e.g. Kvanvig, 2003), in implicitly acknowledging some degree of unity or common ground among the traditionally separate and distinct forms of understanding exemplified e.g. in Windelband's "nomothetic" and "ideographic" modes of inquiry, or the kinds of "understanding" at work in scientific understanding, on the one hand, and such things as "understanding a person" (Grimm, 2016) or "international understanding" (e.g. Grimm, 2018) on the other.

This is not the sort of question to which there could be a "right" answer. There is evidently an "enterprise of understanding," distinct from individual minds, as Howard Stein (2004, p. 135) suggests, entwined with the "enterprise of knowledge" from the beginning, and indispensable to it: "Understanding without knowledge is empty; knowledge without understanding is impossible." Here Stein self-consciously follows in the footsteps of Hermann Weyl, who began his treatise on *Space, Time, Matter* by reminding his readers that the task of understanding refers back to murkier depths than the mathematical methods of scientific knowledge can handle: "Beyond all particular knowledge remains the task of *understanding*. Despite the discouraging spectacle of philosophy careening from system to system, we can't dispense with it if knowledge [*Erkenntnis*] isn't to descend into a pointless chaos" (Weyl, 1918, p. 9). For Weyl, then, the "enterprise" of understanding was something like the philosophical tradition since the ancients. But his own choice of philosophical route to understanding was Husserl's phenomenology, so despite his commitment to an "enterprise" of understanding, he clearly had no qualms about locating understanding in individual subjectivity.

Another reason to relativize understanding, in the first instance, to individuals rather than communities of them, is referred to by de Regt (2017, pp. 37–44; 2020) himself, when he points out that understanding is, after all, at least partly a matter of values. And while there might be a local consensus within some scientific communities about ultimate values – what to *care about*, in other words – any such consensus can never be entirely homogeneous, can hardly be uniform in the weights it assigns to every single argument in any (more or less) shared system of values, let alone in the hierarchy of particular values within that system. This was the preoccupation of Harry Frankfurt (1982), who bit the bullet and saw that what people actually care about – whatever it might have been in a distant, idealized, harmonious past, as imagined e.g. by Alasdair Macintyre (1982) – has essentially become each individual's own personal categorical imperative (Frankfurt 1994). Accordingly there seems to be no natural stopping point in the further disaggregation of de Regt's context of values bearing on understanding until we get to the level of the individual.[1] The remainder of this paper will explore some consequences of that move.

Section 1 will take a step back and very roughly map out the territory before we venture into it. If we are to consider the "individual" as the locus of understanding, what do we mean by that in this context, given the widespread skepticism among cognitive scientists about the very existence or efficacy of such a thing as a human individual "self"? Against this background, Section 2 takes a first stab at isolating and clarifying what could be meant by "objective" and "subjective" content or its understanding in that individual sense. Section 3 will clarify some questions raised by the previous section, and section 4 briefly draws a few provisional conclusions.

1. Who or what does the understanding?

If understanding resides in the individual person, who or what in the person actually does this job? There have been many different candidates – many conceptions of what an individual "self" consists in (see e.g. the survey in Flanagan, 2017, Ch. 11). Since the second half of the twentieth century, however, many traditional conceptions of the soul or self have been called into question by the empirical research of cognitive scientists and neuropsychologists. Beginning perhaps with the specialized, independent "demons" in Oliver Selfridge's 1959 "pandemonium" model of visual perception and learning, cognitive scientists progressively dismantled the more or less unified conceptions of the self, which they disaggregated into specialized, self-regulating "modules" requiring no central supervision. As more and more mental modules were identified and investigated, it was natural to extrapolate this conception to the entire human system of information pooling and seemingly "central" control, to the conclusion that no such control is necessary – so there probably isn't such a thing. Daniel Dennett, who became the best-known philosophical spokesman for this view, maintained that although we perceive a stream of internal consciousness, form

[1] Though that needn't of course mean that, having once got clear about individual subjective understanding, we can't then – that insight in hand – return to social aggregations of individual subjectivities, such as scientific communities, as locations of a collective "enterprise" of understanding.

intentions, and make plans, those are all mere epiphenomena, with no effect on what actually goes on; there is no distinction between what we "do" and what just "happens to us." Our thoughts, feelings, plans and intentions don't influence any of it:

> [Dennett] thinks this running narration is empty chatter: the "I" that is supposed to be the proper subject of experience, thought, and action doesn't exist. In his view, autobiographies are a confabulatory byproduct of the decentralized brain activity that actually regulates behavior: someone who mistakes it for an accurate portrait of the source of behavior makes the same mistake as someone who mistakes the smoke spewed out by an engine as the power that drives it (Ismael, 2006, p. 346).

Moreover, it was discovered experimentally that some actions, at least, are initiated in the nervous system milliseconds *before* their conscious initiation in the mind; the conscious intention is apparently not actually in charge of what we do. At least some of the time, we run on auto-pilot and our conscious intentions are illusory (Libet, 2004).

Such findings, and others like them, elicited varied responses. Some go along with Dennett, and doubt that there is a self that makes a difference (Metzinger, 2009; Olson, 1999). Others, however, have resisted this conclusion, for a wide variety of reasons, which we can group – oversimplifying somewhat – under three broad headings: First, those who stick with our native intuitions of autonomy and conclude that there must therefore be something radically incomplete about our entire scientific edifice, perhaps with physics itself. This position, associated with David Chalmers (1996), revives something like Cartesian dualism. Second, those who largely adhere to our basic science but are critical of some parts of current biology for its over-reliance on a particular evolutionary paradigm inherited from Darwin and propagated through Dawkins and Dennett (Thompson, 2007).[2] Third and finally, there are those who adhere completely (or almost completely) to current science, including biology, but who see the cognitive-science reduction of the self to self-organizing systems as an unnecessary and unwarranted extrapolation from the available empirical findings, and who find room for a top meta-layer of a *self-governing* system that may not be as fully autonomous as our intuitions (or Descartes) make it seem to us, but is able to plan, coordinate, and execute our daily lives to some significant degree. Notable representatives of this third position are, especially, the computer scientist Douglas Hofstaedter (2007), the neurophysiologist Merlin Donald (2001), and the philosopher of physics Jenann Ismael (2007, 2010, 2016).

While all three responses have their followings, and plausible arguments can be advanced for all three, the revival of dualism has lost some of its initial appeal and

[2]Adherents of the "predictive processing" program, motivated by Friston's "free energy principle" (FEP) have also tended to this second view; the most explicitly elaborated account of the "self" in this framework is Kiverstein (2020), which tries to incorporate significant elements of Varela's and Thompson's "enactive" approach, and endorses a version of "autopoiesis" indebted to it. However, its adherents Di Paolo, Thompson, and Beer (2022) argue that this attempt fails, as the FEP is fundamentally incompatible with their view. It is not even clear (e.g. Klein 2021) that anything like a distinction between beliefs and desires (even of the minimal sort relied on in section 2 below) is possible within the FEP framework.

shock value in the face of some effective refutations (Perry, 2001; Ismael, 2007, Ch. 8-10), which have mostly gone unanswered. The second position has its attractions, but regarding the main issues at stake for understanding, there is little difference between the second and third positions. Since the third position requires fewer adjustments to our overall scientific conception (and is in that sense more conservative and less disruptive), this paper will base its further discussion of what goes on when an individual understands something on this third conception of the self.

This third view takes on board the cognitive science, modularist, self-regulative dissolution of the intuitive self but nonetheless arrives at a certain vindication of our instinctive beliefs about our ability to exert at least a degree of control over our lives, plans, and actions. Ismael develops this idea further to argue for the genuine (if limited) autonomy of the self, so conceived, from traditional worries about determinism (Ismael, 2016). In the resulting picture, the self is no longer a *thing*, it isn't located anywhere in the brain; it is a viewpoint, a self-locating representation within its (physical and social) environment as well as among the many self-organizing processes inside the brain. It is a fixed point at which its meta-viewpoint on those processes can be coordinated with those processes themselves and the environment(s) they inform about. Dennett (1992) had himself conceived of the self similarly, as a "center of narrative gravity" that begins as a helpless plaything of the self-organizing processes going on around it but gradually accumulates a continuous memory of its efforts to assert itself in that chaotic struggle:

> [T]he extra level of explicit representation is what opens up the space for the formation of an internal point of view. At first, this internal point of view is nothing more than the self-locating "I" that identifies a vantage point on space, but as one moves through the world, acting and observing the results of one's actions, building up an internal storehouse of historical memories, a process begins to take hold that supports the development of dispositions, preferences, tastes, and values. The brain builds up a library of ideas that encode information about parts of the environment that form natural dynamical units – complex systems with enough internal integrity to allow them to be tracked over time. This library of ideas forms the basis for a set of concepts that are given names and become the components of Fregean thoughts and vehicles of interpersonal communication. The early glimmers of self-hood eventually turn into a fully-fledged sense of oneself as a locus of value, social agent, possessor of rights and obligations, a partner in social relationships, and so on …all the trappings of personhood. (Ismael, 2011, p. 739)

To Dennett, however, this emergent center of narrative gravity is just that – it tells the tale, but plays no significant part in the action, and even the tale is only for external, social presentation; it bears at most a tenuous relation to any actual history and has no effect on it.1 It is just a self-referential feedback loop, a fixed point at which the meta-level meets the object level of self-organizing modules, a "strange loop," as Hofstaedter calls it. Ismael agrees that it can't pretend to the sovereign dictatorship over its fate that the traditional self or "soul" imagined it could wield, but she nonetheless advances a multi-pronged battery of arguments to establish a

significant role for a more modestly conceived self, on the model of the executive committee of a large, relatively immobile corporation. This is a rather insubstantial, evanescent self compared to that of Descartes, all but invisible from its own viewpoint; it can detect only the *objects* of its attention, not itself. (Hume was right, in that respect!) It can allocate attention, though its attention (working memory) capacity is notoriously very limited. But there is a crucial difference, she argues, between mere self-organization and self-*governance* (Ismael, 2016, Ch. 2); self-governing systems are those "in which at least some organized activity is the result of a centralized process that involves the sharing of information and the formation of an overall plan and deliberate coordination of activity" (ibid., p. 19):

> Self-governance contrasts with pure self-organization. In a purely self-organizing system, all behavior is emergent from the aggregated activity of components, each doing its own thing. The coupling among components can generate the appearance of coordination, but there is not really any pooling of information and centralized control of activity. In a self-governing system, by contrast, at least some of the information distributed throughout the system is collected, synthesized, and used to fuel a decision procedure that plays a role in guiding the system's behavior. The decision procedure not only draws on information that is distributed throughout the system, but It can coordinate the activity of spatially separated components. (ibid., p. 20)

The CEO is ultimately responsible, but doesn't exercise control by fiat – by pressing a button or just issuing commands. Her efficacy is submerged in that of an executive committee who have to be persuaded or cajoled into an agreed common policy (requiring compromise on her part as well as theirs), which then has to be implemented over all the obstacles presented by the inertia of a large, hidebound organization full of self-organizing units that may or may not not be amenable to new initiatives from above.

The "internal storehouse of historical memories" gradually acquires a structure, becoming something like what is often assumed in the background in cognitive science (and called an internal "knowledge base" or "world model" (or "mental life").[3] So it isn't just a storehouse, it has an internal *organization* – beginning with a default organization inherited from the child's immediate linguistic surroundings (an organization built in to the way the language is used in the childhood surroundings, the socialization

[3]The "knowledge base" was the backbone of the "expert systems" that dominated AI research in the 1980s; "world models" appear now to be making a comeback, as the increasingly apparent shortcomings of LLMs have convinced some leading AI researchers that a world model needs to be supplied as a relatively fixed background and can't be relied on to arise spontaneously as a by-product of deep learning, though this is still controversial. The cognitive-science terminology varies; Kintsch (1988) called a local version (relevant to a particular text) a "situation model"; Bereiter (1990) called a more extensive version a "contextual module" and later an even more comprehensive version (Bereiter and Scardamalia, 2002) a "mental life." The latter term is favored here since it is applied somewhat more broadly (though still not, as in the present paper, including values) than is typical in cognitive science or AI, where it is largely restricted to knowledge or the purely cognitive. See also the discussion at the beginning of section 2.

of origin). The "process" that "begins to take hold" isn't automatic; it requires a certain amount of prompting from those immediate surroundings, or a reaction against them. And it's not this process itself that (unilaterally or automatically) gives rise to the "dispositions, preferences, tastes, and values"; the process is kicked into motion by some endowment inherent in the particular individual's dispositions that initiates a *dialectic* between those predispositions and the progressive accumulation of experiences of interacting with the immediate (and then wider) environment. The process is neither automatic nor passive; it is initiated by the incipient person and is from that first moment a dialectical one, a mutual feedback relation between the person and her environment.

At some point in many people's early development, in childhood or youth, the system of "dispositions, preferences, tastes" gradually becomes less miscellaneous and more structured and hierarchical; it begins to be organized by incipient *values*, whose objects (things, experiences, feelings, people) the person finds she *cares about*. Here we have arrived in Frankfurt's (1982, 1999, 2006) territory; the objects of the values-in-the-making are the things, people, attitudes that, more and more, come to *matter* to her. She can, of course, be disappointed in some provisional objects of these values, and may find that she really didn't value that person or feeling or thing as highly as she first imagined. Or the value itself remains intact but the tentatively chosen *object* is found to fall short of the incipient value formed in anticipatory appreciation of it. This trial-and-error process can lead, then, to the values themselves (as distinct from just their objects) rising to the level of conscious awareness (which does not mean that they can be articulated in words).

What is built up as a result of such a process over the years – Ismael's "library of ideas" – gradually becomes even more hierarchical as the incipient values compete with each other – especially if and when they have risen to conscious awareness – for primacy within a provisional value system or hierarchy (which again need not be articulated in literal language). So what is built up is not so much a "library" as a hierarchical *system* that maintains strong continuities over time but is constantly under revision at the margins, in response to new episodes of *understanding*, i.e. of fitting new encounters with new potential objects (sought out or involuntarily thrust upon one) into the existing system, the existing mental life, which is thereby modified.

The central protagonist in such processes of understanding is not so much the shadowy, flickering subject of attention, the "strange loop" itself and its limited working memory, as the mental life, the lifelong construction of an inner world, a system of values within a progressively more coherent representation of its natural and social environments. Understanding can be seen as the process of investment in this inner world, of confronting it with new inputs, and its (eventual) transformation in response. Though the minimal, evanescent, situated self is highly constrained, and its capacity very limited, its one crucial capability is its ability to focus (mentally "top-down") *attention* on both the world around us (to update on it continuously) as well as on that inner world, and to coordinate them. This capability of allocating attention acts as the interface between the minimal self and its accumulated mental life in both directions: Attention seeks out and selects the inputs relevant to the existing mental

life and integrates them with it by discerning where and how they fit (or don't). It also has the essential function, in the other direction, of supporting the evanescent shadow of a self with a scaffolding that keeps it upright and on track. Without such a supporting structure, its limited capacity could hardly keep in mind who it is and what it cares about from hour to hour. It has to remind itself constantly of these basics – by allocating attention to them – to supply itself with substance or continuity ("character"). Attention to the accumulated mental life "steadies the mind," as Bernard Williams (2002, pp. 185–98) put it, by orienting it and reminding it "who it is" (and where that comes from).[4] Carolyn Dicey Jennings (2020) seeks to establish (by extensive empirical as well as philosophical arguments) that a metaphysically substantial self must be the subject of (top-down) attention, but even the rather ghostly, insubstantial self suggested by Ismael and Hofstaedter can fulfil the key self-governing function of allocating attention by retrieving what it needs (i.e. reminding itself what it thinks and who it is) from the rich mental life it has invested in over a lifetime.

2. Objective and subjective understanding

The "world model" or "mental life" assumed in the background by cognitive scientists and in 1980s-90s AI (see above, footnote 4) is usually treated as a *knowledge* base, like Schurz and Lambert's "cognitive background corpus," abstracted from any affective or conative element. This assumed background setting in cognitive science is generally not intended as realistic; it serves merely as a makeshift background prop for a particular investigation. When we take it seriously as an actual substrate of (and repository for) a human self, though, it turns out that the affective and conative components of human experience and deliberation are impossible to extricate from the cognitive component, as James Woodward (2016) has convincingly argued, on the basis of more recent cognitive science. So as a background to the workings of understanding, the "mental life" into which a new experience or phenomenon is integrated is here conceived as holistic, including not just knowledge and cognition but also affect, sensory experience, desires, moods, values, and whatever else is present to mind. It is not just a *world* model, but includes also our responses to the world and what we *want* from it. (So the "knowledge" itself, in this "knowledge base," includes not just formal knowledge, but all the informal, tacit, contextual knowledge by which people find their way into formal knowledge and get comfortable with it so as learn where it is relevant and how to use it; e.g. Tall, 2013.)

We initially analyze this comprehensive mental life into two components: beliefs and desires, crudely speaking. Or, to borrow a somewhat subtler conceptual tool from micro-economics, we call them *constraint structure* and *optimand structure* (or *value system*). The constraint structure corresponds roughly to the purely cognitive world model employed in older cognitive science and AI, as just discussed, and was

[4] Thomas Mann once said (in English, at Princeton), "One always needs to be reminded; one is by no means always in possession of one's whole self. Our consciousness is feeble; only in moments of unusual clarity and vision do we really know about ourselves." (Mann, 1953, p. 45)

then often highly oversimplified.[5] In humans it is the "real world" – the constraints subject to which the self optimizes its pursuit of the things it cares about (its values). The constraint structure is the same for everyone, even if the aspects of it that are relevant to individual decisions and life plans are different for each person (since each person has different values, subject to different constraints in the world out there). And large parts of the constraint structure (of "reality") *are* relevant to everyone, e.g. the knowledge that we all die after living a certain number of decades, that illness and accidents threaten life and the quality of life, that nourishment and shelter are conducive to survival, and other such universally relevant basics.

Correspondingly, individual value systems, however they ultimately vary, nearly all include certain basics in common: one values one's own survival, avoidance of pain and illness, survival of one's friends and family, and so on. These are so widely shared that Frankfurt (2006, pp. 34-38) suggests that we regard their presence in a value system as a criterion of "rationality."[6] He also assumes that by virtue of a (loosely defined) "practical reason" nearly everyone develops secondary, derived values as practical consequences of, or practical prerequisites for, the primary things they care about (ibid., p. 20), so that their value system comes to include not only the primary values (what they care about directly) but also a much wider infrastructure of civic responsibilities, institutional loyalties, and moral principles.

The *weights* of the widely-shared basic values within a value system, however, change with circumstances. Ordinarily, most people devote little conscious thought or affect to the avoidance of death; only when life is threatened does survival itself suddenly become a major factor in what one cares about. So the weights of the values in one's value system can vary with situations and with time. The constraint structure, in contrast, is comparatively stable; the "real world" doesn't change much from day to day or year to year. We can assume a basic asymmetry, then, between the constraint structure and the value system.

Most people's knowledge of the constraint structure doesn't extend very far. In most first-world countries, it does, however, include rudimentary versions of relatively recent scientific knowledge: day and night are due to the earth's rotation on its axis, infectious diseases are caused by micro-organisms and viruses, heredity is encoded in a complicated molecule called DNA, humans are descended of primates similar to present-day chimpanzees, and so on. A small minority are voluntarily encultured into more sophisticated versions, in which they come to appreciate that such widely-shared fragments of knowledge fit together (quite apart from any supposed reducibility or "unity of science") into a more or less coherent scientific world view.[7] This world view

[5] As in Kintsch (1988), where it consists only in the background knowledge required to understand a particular text. An extreme case of such oversimplification in early AI days was Terry Winograd's SHRDLU, whose "world" contained a total of only 50 items and operations on them.

[6] And on that basis challenges Hume's notorious claim that it can be "rational" (not contrary to "reason") to prefer the destruction of the human race to a scratch on my finger (ibid., pp. 29-30).

[7] Which is not to suggest that the process of understanding be *defined* as tending by default toward increased unification (of the agent's entire "cognitive background corpus"), as Schurz and Lambert (1994) do. Their conception is highly idealized; a psychologically and sociologically more realistic account would need to consider all the empirical obstacles individual subjectivities face to the genuine,

is something like a canonical description of our shared human constraint structure; it "limns the contours" of the reality we all face as human beings and the limits of what we can aspire to as inhabitants of that reality. To that extent, the constraint structure is "objective" while the value system – though parts of it are shared and may count as basic or rational – is comparatively "subjective."[8]

To understand an aspect or component of the constraint structure – to incorporate it into one's mental life – is to understand something objective, then, in that sense. But even such objective understanding is a matter of degree. One can understand a new fact or concept one hadn't encountered before without realizing many of its implications; these may gradually sink in as one thinks about it (allocates attention to it) in different contexts. One can understand something in the abstract without being able to apply it very flexibly; again, it may or may not sink in as one begins to try to use it for different purposes and to embed it in a corpus of informal and situated knowledge, find a home for it, so to speak, among the other facts and theories one may have domesticated in one's mental life over the years.[9] As these implications and uses and extensions become more habitual and familiar, the newly understood item has an impact on the entire mental life it was incorporated into; one might say that the new item isn't "fully" understood until one's mental life has completely accommodated it in all these ways.[10]

Understanding values, or something relevant to values, is analogous. One can incorporate a new value or viewpoint into one's value system, and since values – what someone cares about – as discussed here (following Frankfurt) are subjective, we call this subjective understanding. As in the case of objective understanding, such incorporation is a matter of degree. A new understanding eventually or potentially has an impact on the whole value system, but takes time and attention to sink in, and has to find a home in surroundings that may at first seem alien. Understandings assimilated into a value system can vary widely; they can be anything with a potential impact on (relevance to) what a person cares about. De Regt's (2025, pp. 31–34) music example is an instance of subjective understanding in this sense. As he implies, a similar dynamic or capability of the human mind is involved as in scientific understanding

objective unification Schurz and Lambert primarily consider (though it is defined more broadly than the merely deductive unification originally proposed as the criterion of understanding in Friedman, 1974).

[8]Kantians do not, of course, recognize this distinction; for them the categorical imperative is just as objective as the constraint structure.

[9]Cf. *Nicomachean Ethics* Bk. 7, Ch. 3, where Aristotle blames akrasia on superficial knowledge (of what one knows to be the right thing) as opposed to "knowledge proper," by which he evidently means something like "full understanding."

[10]This is, of course, a somewhat oversimplified account. One important dimension it omits is the "division of linguistic labor" as applied specifically to understanding: in some cases we purposely – strategically – leave the "full" understanding of certain specialized concepts or ideas to specialists and hope we've understood enough of the "gist" to serve our own purposes (see Keil, 2019 on this sort of purposeful "partial understanding"). But such offloading of cognitive responsibility requires a degree of metacognitive awareness of the limits of one's own understanding, unlike the partial understanding of the beginning student who still struggles to fit a new item into her mental life but is unsure where it might fit (see also Tall, 2013).

(i.e. understanding of the constraint structure). In both cases there is an analogous incorporation of a new element into an existing framework. The existing framework is not a *knowledge* framework in this case, though, it is a *values* framework.[11] When a new value element has been assimilated into it, it changes that framework, which adjusts and reorganizes to accommodate the new arrival.

Though the processing of the piece of music, resulting gradually in its understanding, in de Regt's example, has both objective and subjective components, the content of the piece (not discussed by de Regt) is something – an ideal, a gesture, a way of being – that one can recognize, subjectively, as aligned or contrary (or oblique) to what one cares about, i.e. to one's own values.[12] So the understanding is subjective; whether something counts as objective or subjective understanding, as these terms are used here, depends on the *content* being processed, not on the component of the mental life (constraint structure or value system) affected by that content. Objective content concerns the constraint structure; subjective content concerns the value system. They interact; objective understanding impacts the value system; understanding that mRNA vaccines do not alter the genes of the person vaccinated can, for instance, influence what an anti-vax fanatic cares about. Likewise subjective understanding influences a person's priorities about which components of the constraint structure matter to her. Each instance of both objective and subjective understanding affects the assimilation of a new item into the *whole* mental life.[13]

3. Some clarifications

It is tempting to try to assimilate this view of understanding into other psychological (and/or moral-psychology) frameworks. For those acquainted with dual-process conceptions of human mental processing,[14] for instance, it may seem at first glance as if objective understanding is the province of what Kahneman (2011) calls "system 2" while subjective understanding is a matter of his "system 1." But as we saw in de Regt's music example, there are often "system 2" aspects in the understanding of subjective

[11] Not that these can of course be tidily separated in actual fact; the constraint structure constrains the pursuit of values and the values determine which parts of the constraint structure are relevant to it (are sought out for attention). They interact. But the value system and the constraint structure (the conative and the cognitive) can still be analytically distinguished as components of a mental life, just as consumers and producers are distinguished in economics although nearly everyone is in fact both.

[12] Taylor (2010, p. 35) focuses on the actual content of a musical piece, to motivate a distinction between "articulative" and "disclosive" modes of communication (or uses of language). However one conceives of this distinction, it does not undermine de Regt's thesis (endorsed here) that subjective (e.g. musical) and objective (scientific) understanding are rooted in the same (or a very similar) human mental capability.

[13] An anonymous reviewer pointed out that the understanding of *persons* involves even more complex interactions between objective and subjective understanding, as those terms are here understood. Grimm (2016) suggests that objective understanding of the other person's actions or emotional states can be a precondition for the (further) subjective understanding of their values. For a more complete picture of objective and subjective understanding, as here intended, such complications would need to be pursued in more detail than this paper allows.

[14] As widely popularized by Stanovich (2004) and Kahneman (2011).

(values-oriented) content, while *any* employment of system 2 requires the underlying (system-1-) automaticity needed for the use of any conceptual system (e.g. the use of a spoken language requires automaticity in recognizing that language's phonemes, the use of written language requires automaticity in reading, etc.). So the distinction between objective and subjective understanding is oblique to the distinction between system 1 and system 2.

System 1 is often seen (by Kahneman himself and many others) as at best partly conscious, which again suggests that perhaps subjective understanding reaches into murky intuitive (subconscious) depths while objective understanding is a matter of fully explicit consciousness. But subjective (value-system-oriented) understanding is not restricted to the affective or the instinctive; the extent and complexity of conscious deliberation that goes not just into the creation but also the reception and the understanding of a work of art such as a poem is well illustrated by Matthew Hollis's (2022) "biography" of T.S. Eliot's *The Waste Land*. Conversely, the subconscious roots of scientific creativity (about which as little is known as about any other sort of creativity) have been widely discussed. So once again, the distinction between objective and subjective understanding is oblique to that between conscious and unconscious processing.

A third (and related) potential misconstrual of the distinction between objective and subjective understanding is to restrict objective understanding to the information-conveying or "articulative" (or "assertive") mode of communication that Charles Taylor contrasts with the "disclosive" mode (footnote 13 above), and to regard the latter as the only possible medium for subjective understanding. But in de Regt's example, a piece of music is subject to both objective and subjective understanding, and both contribute to understanding the subjective (value-relevant) content of the music. So "articulative" analysis of harmonic and contrapuntal relations can support subjective understanding, and the restriction of objective understanding to "articulative" communication (or subjective understanding to "disclosive" communication) breaks down.

4. Further steps

If we take seriously the idea that understanding is, in the first instance, located in individual subjectivity (even if we then focus our interest, like de Regt, on communities of individuals with similar training and enculturation), this paper has argued, then the individual's mental life (comprehensive world model) is the primary venue of understanding. The self (in the sense of Ismael or Hofstaedter) can't understand on its own, it can only allocate attention to the incorporation of a new encounter into its mental life (which supplies the self with its substance or content), but understanding is hardly complete at the moment of this first encounter; continued investment of attention is required for the process to go on and reach some degree of completion. As the new item to be understood sinks in, over longer periods, its implications – logical as well as practical – gradually become evident. The "mental life" into which the new encounter is incorporated is conceived as a hierarchically structured subset of an individual's total memory in which values (what the person cares about, in Frankfurt's

sense) govern priorities both in the realm of action and in the realm of updating/getting informed about the constraint structure (the "real world").

This paper has further suggested, in agreement with de Regt, that objective (constraint-structure-oriented) and subjective (values-oriented) understanding proceed from the same mental capacity or function and differ only in the content (concerning either constraint structure or value system) they respectively process. In both cases, understanding consists in the assimilation of a new experience or encounter into an existing mental life. As it progressively assimilates new encounters, the mental life is in constant flux. Sometimes the assimilation is frictionless; the new element fits well with what is already there. More often, there will be friction, but there is no way to know a priori where, in what parts of the mental life (overall world model), the eventual long-term adjustments will turn out to be, and what if anything will change as a result.[15]

Is this conception, or are any parts of it, empirically testable? In principle yes, but unfortunately "world models" have themselves hardly been a focus of research in cognitive science, especially since expert systems went out of fashion in AI. In the study of human cognition, world models were usually rudimentary place-holders, and even where they came into focus, everything but propositional knowledge was abstracted from them (e.g. Bereiter and Scardamalia, 2002), whereas it is essential to the proposal suggested here that the mental life be conceived as *including* affective, sensory, and conative components. Another complication is that mental lives are long-term processes that would have to be studied longitudinally, to follow and describe the longer-term process of understanding – i.e. of investment in the assimilation of encounters with new elements into the mental life. Longitudinal research is not only expensive and unwieldy, since research subjects have to be kept track of and re-tested and re-interviewed at intervals for years. (It also, therefore, doesn't lend itself to the typical time-frame of Ph.D. research.)

Before understanding (the assimilation of a new item into a mental life) can be empirically studied, though, mental lives themselves need to be better understood. We don't even have anything like a basic geography of the territory. Clearly something like Searle's (1983, Ch. 6) "background" must underlie an individual's mental life (perhaps in the holistic form suggested by Dreyfus, 2012, following Merleau-Ponty), and the relation between this (submerged or subliminally conscious) background and the mental life (which has been conceived here as accessible to consciousness and selectively retrievable more or less at will) could be investigated empirically, but this would have to be preceded by an inventory of a mental life's contents. Even a rudimentary catalogue of the basic substrate of "commonsense knowledge" underlying any individual's conception of the constraint structure (such as that initiated by Ernest Davis 1990 and later pursued at an industrial scale by Douglas Lenat, e.g. 1995) is a staggeringly vast undertaking. We also do not even have the beginnings of a framework in which to characterize the parameters along which individual minds vary in their

[15] So there is room for considerable "freedom of understanding" of the kind discussed (and advocated from a very different perspective from that of this paper) in Bertram (2024), who seeks a reconciliation between hermeneutics and "critical theory," traditionally seen as opposed, on this basis.

capacities or even their basic structure (Strawson, 2009, pp. 15–18). And so on; just to list the extent of our ignorance could easily take many pages. Nor do we even really know how to approach these questions; while there is undoubtedly a role for the traditionally standard toolbox of psychometric and neuropsychological modes of inquiry, some now think that these need to be supplemented or at least partly replaced by a "new science of consciousness," empirical studies of the subjective world within (e.g. Thompson, 2017).

Meanwhile, both phenomenological explorations and extrapolative suggestions from what we do know, such as those in this paper, can help to guide future empirical research, either by indicating how a schematic framework for such research could look, or by eliciting contradiction that could lead to better and more empirically fruitful frameworks.

References

Baumberger, C., Beisbart, C., & Brun, G. (2017). What is understanding? An overview of recent debates in epistemology and philosophy of science. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 1–34). Abingdon.

Bereiter, C. (1990). Aspects of an educational learning theory. *Review of Educational Research, 60*, 603–624.

Bereiter, C., & Scardamalia, M. (2002). Schooling and the growth of intentional cognition: Helping children take charge of their own minds. In B. Smith (Ed.), *Liberal education in a knowledge society* (pp. 245–277). LaSalle, IL.

Bertram, G. W. (2024). *Die Freiheit des Verstehens: Eine hermeneutisch-kritische Theorie*. Frankfurt/Main.

Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Davis, E. (1990). *Representations of commonsense knowledge*. San Mateo, CA.

Dennett, D. C. (1992). The self as a center of narrative gravity. In F. S. Kessel, P. M. Cole, & D. L. Johnson (Eds.), *Self and consciousness: Multiple perspectives* (pp. 103–115). Hillsdale, NJ.

de Regt, H. W. (2017). *Understanding scientific understanding*. Oxford University Press.

de Regt, H. W. (2020). Understanding, values, and the aims of science. *Philosophy of Science, 87*, 921–932.

de Regt, H. W. (2025). Understanding in science and beyond. In A. I. Mărășoiu & M. Dumitru (Eds.), *Understanding and conscious experience: Philosophical and scientific perspectives* (pp. 23-41). Abingdon.

Di Paolo, E., Thompson, E., & Beer, R. D. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences, 3*, 1–39.

Donald, M. (2001). *A mind so rare: The evolution of human consciousness*. New York.

Dreyfus, H. (2012). Introductory essay: The mystery of the background qua background. In Z. Radman (Ed.), *Knowing without thinking: Mind, action, cognition, and the phenomenon of the background* (pp. 1–10). Basingstoke.

Flanagan, O. (2017). *The geography of morals: Varieties of moral possibility*. Oxford.

Frankfurt, H. (1982). The importance of what we care about. *Synthese, 53*, 257–272.

Frankfurt, H. G. (1994). Autonomy, necessity, and love. In H. F. Fulda & R.-P. Horstmann (Eds.), *Vernunftbegriffe in der Moderne: Stuttgarter Hegel-Kongress 1993* (pp. 433–447). Stuttgart. (Reprinted in H. G. Frankfurt, *Necessity, volition, and love* (pp. 129–141). Cambridge.)

Frankfurt, H. G. (1999). On caring. In H. G. Frankfurt, *Necessity, volition, and love* (pp. 155–180). Cambridge.

Frankfurt, H. G. (2004). *Taking ourselves seriously and getting it right*. Stanford, CA.

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy, 71*, 5–19.

Grimm, S. R. (2016). How understanding people differs from understanding the natural world. *Philosophical Issues, 26*, 209–225.

Grimm, S. R. (2018). The ethics of understanding. In S. R. Grimm (Ed.), *Making sense of the world: New essays on the philosophy of understanding* (pp. 116–134). Oxford.

Hollis, M. (2022). *The Waste Land: A biography of a poem*. London.

Ismael, J. T. (2006). Saving the baby: Dennett on autobiography, agency, and the self. *Philosophical Psychology, 19*, 345–360.

Ismael, J. T. (2007). *The situated self*. Oxford.

Ismael, J. T. (2010). Me, again. In J. K. Campbell, M. O'Rourke, & H. S. Silverstein (Eds.), *Time and identity* (pp. 209–227). Cambridge, MA.

Ismael, J. T. (2011). Précis of *The situated self. Philosophy and Phenomenological Research, 82*, 733–758.

Ismael, J. T. (2016). *How physics makes us free*. Oxford.

Jennings, C. D. (n.d.). *The attending mind*. Cambridge.

Kahneman, D. (2011). *Thinking, fast and slow*. New York.

Keil, F. (2019). How do partial understandings work? In S. R. Grimm (Ed.), *Varieties of understanding: New perspectives from philosophy, psychology, and theology* (pp. 191–208). Oxford.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95*, 163–182.

Kiverstein, J. (2020). Free energy and the self: An ecological–enactive interpretation. *Topoi, 39*, 559–574.

Klein, C. (2021). A Humean challenge to predictive coding. In D. Mendonça, M. Curado, & S. S. Gouveia (Eds.), The philosophy and science of predictive processing (pp. 25–38). London.

Kvanvig, J. (2003). *The value of knowledge and the pursuit of understanding*. Cambridge.

Lenat, D. (1995). A large-scale investment in knowledge infrastructure. *Communications of the ACM, 38*(11), 33–38.

Libet, B. (2004). *Mind time: The temporal factor in consciousness*. Cambridge, MA.

MacIntyre, A. (1982). Comments on Frankfurt. *Synthese, 53*, 291–294.

Metzinger, T. (2009). *The ego tunnel: The science of the mind and the myth of the self*. New York.

Olson, E. T. (1999). There is no problem of the self. In S. Gallagher & J. Shear (Eds.), *Models of the self* (pp. 49–61). Thorverton.

Perry, J. (2001). *Knowledge, possibility, and consciousness*. Cambridge, MA.

Schurz, G., & Lambert, K. (1994). Outlines of a theory of scientific understanding. *Synthese, 101*, 65-120.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge.

Stanovich, K. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago.

Stein, H. (2004). The enterprise of understanding and the enterprise of knowledge. *Synthese, 140*, 135–176.

Strawson, G. (2009). *Selves: An essay in revisionary metaphysics*. Oxford.

Tall, D. (2013). *How humans learn to think mathematically: Exploring the three worlds of mathematics*. Cambridge.

Taylor, C. (2010). Language not mysterious? In B. Weiss & J. Wanderer (Eds.), *Reading Brandom: On making it explicit* (pp. 32–47). London.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of the mind*. Cambridge, MA.

Thompson, E. (2017). *Waking, dreaming, being: Self and consciousness in neuroscience, meditation, and philosophy*. New York.

Weyl, H. (1918). *Raum–Zeit–Materie: Vorlesungen über allgemeine Relativitätstheorie*. Berlin.

Woodward, J. (2016). Emotion vs. cognition in moral decision-making: A dubious dichotomy. In S. M. Liao (Ed.), *Moral brains: The neuroscience of morality* (pp. 87–116). Oxford.

Balkan Journal of Philosophy (2026), vol. 18, issue 1, 66 – 85

ON TWO ACCOUNTS OF MISUNDERSTANDING: A REPLY TO COLLIN RICE AND KAREEM KHALIFA

Stefan Petkov*

Abstract: When developing theories of understanding, philosophers have tried to provide a full package by offering criteria both for the epistemic success of understanding and for various related epistemic failures. One of those failures is misunderstanding. Currently, there are only two theories of misunderstanding. The first one, from Yu and Petkov, analyzes the unique epistemic conditions for misunderstanding. The other one, by Rice and Khalifa, studies the process of correcting misunderstanding. In this paper, I show that both theories have shortcomings. I argue that Rice and Khalifa's theory is too broad and permissive. It effectively blurs existing distinctions between proto-understanding or understanding based on falsified theories. The reverse holds for Yu and Petkov's theory. It is too narrow and thus fails to capture broad issues related to clearing misunderstanding. I conclude by suggesting that Yu and Petkov's theory describes misunderstanding more plausibly, whilst Rice and Khalifa's theory can capture the correction of misunderstanding.

Keywords: Explanatory understanding, misunderstanding, epistemology of science, disputes in science and philosophy.

1. Introduction

The philosophical analysis of understanding has shown that it is an epistemic category that captures a rich spectrum of phenomena (Baumberger, 2019; Lawler et al., 2023). Philosophers typically provide a unified framework of this spectrum, by focusing on the criteria for the epistemic success of understanding, whilst expecting that the negative epistemic outcomes would naturally follow, as a contrast. Recent interest in misunderstanding, as a specific epistemic error, has shown that this is not necessarily the case (Petkov & Yu, 2023; Yu & Petkov, 2024; Rice & Khalifa, 2025). According to these authors, a unified theory of understanding needs to encompass the specific conditions under which both understanding and the negative outcomes occur, such as lack of understanding and misunderstanding. These discussions have further shown that misunderstanding can both be a specific epistemic failure, and that it relates to unique corrective epistemic processes. We can then raise two questions:

(1) Are these theories internally coherent, by keeping intact existing epistemic distinctions?

(2) Are these theories rich enough to capture all related phenomena?

*Institute of Philosophy and Sociology, Bulgarian Academy of Sciences; yaggdrasil@yahoo.com

To answer these questions, I shall argue that Rice and Khalifa[1]'s theory is too broad, whilst Yu and Petkov[2]'s theory is too narrow. Their shortcomings all follow from their scope. RK's definition of misunderstanding effectively blurs existing distinctions made by Khalifa himself – the distinctions between proto-understanding, and understanding based on explanations derived from falsified theories. YP's definition focuses narrowly on misunderstanding as the epistemic error that occurs when a correct explanation is rejected. Yet it fails to properly deliberate between all the consequences that follow from misunderstanding (as they define it).

Nevertheless, I believe that YP's characterization of misunderstanding outperforms its rival. For, not only does it describe misunderstanding more plausibly, but it keeps existing epistemic distinctions. RK's theory, on the other hand, still offers an interesting analysis on how misunderstanding can be corrected. Both theories then can be combined.

To argue for these points, I shall first provide a short sketch of what understanding is according to both theories (section 2). Then, I will outline the two theories of misunderstanding and their chief goals (section 3). In section 4, I shall provide criticisms of both theories and defend the idea that YP's theory remains a better characterization of misunderstanding, whilst RK's theory can still handle how misunderstanding is corrected. I shall conclude the paper by briefly attempting to combine both theories (section 5).

2. Explanatory understanding

Both theories of misunderstanding are based on narrow definitions of understanding, as the epistemic success that results from a correct explanation. Since my goal here is not to extend these theories, I shall follow this narrow context.

I begin with the definitions of understanding embraced by both RK and YP. This step is important, because any theory of understanding should be internally coherent. If a new phenomenon is introduced within a theory, it should not happen at the cost of blurring existing distinctions that the theory has already made. Thus, a definition of misunderstanding should be sufficiently robust. To recognize whether this is something that both theories achieve, we should first outline their existing definitions of understanding, and then check if any of these definitions overlap with the newly introduced category of misunderstanding. If there is such an overlap within some theory, then this theory loses internal coherence.

RK do not offer an explicit definition of understanding, but I assume that Khalifa follows his earlier work, in which minimal understanding[3] is defined as follows:

[1] Abbreviated as RK henceforth.

[2] Abbreviated as YP henceforth.

[3] It is understanding based on a singular instance of an explanation, as opposed to understanding based on a nexus of complementing explanations for the same fact.

> S has minimal understanding of why *p* if and only if, for some *q*, ***S believes that*** *q explains why p, and q explains why p* ***is approximately true***. (Khalifa, 2017, p. 55)

For YP, understanding is defined as:

> the epistemic achievement that results from ***(subjectively) grasping or constructing*** the ***(objectively) correct explanation*** for a given fact (Yu & Petkov, 2024, p. 4).

As we can see both theories differ in the requirements for epistemic subjects and for explanations themselves. Following YP, let's label these criteria or requirements, *subjective* for the former, and *objective* for the latter.

2.1. The subjective criteria for understanding

For the subjective criteria, Khalifa defends the received view that understanding is a type of knowledge – knowledge of a true explanation (Khalifa, 2017). There are only two paths by which such knowledge can be obtained – by testimony and by direct acquaintance. When one discovers a novel explanation, one relies on her cognitive abilities, skills and previous knowledge to reach new explanatory information and thus obtains knowledge by direct acquaintance. When one understands something by testimony, she is offered an explanation that she believes. Khalifa further describes believing as proficiency of using the concepts involved (Khalifa, 2017; 2023). This proficiency for him naturally involves abilities and skills.

Effectively HP's theory offers almost the same criteria. Terminologically they define the criterion for accepting a ready-made explanation as *grasping*, and the criteria for developing a new explanation as *constructing*. For them, grasping similarly designates the abilities and skills to "assimilate" explanatory information. They explicitly state that the relevant cognitive abilities and skills for grasping could be the same as obtaining knowledge. Thus, relevant to the role of skills and abilities, their view is consistent with Khalifa's received view. The case for constructing a novel explanation follows the same logic.

However, YP emphasize that what distinguishes between knowledge and understanding is not the relevant abilities and skills. Rather, it is that understanding occurs as an epistemic gain after completing a specific cognitive task – that of seeking an explanation and asserting that some explanation offers the right type of information.

In the case of grasping an explanation, the subject *sees*[4] that the information being offered solves the problem. In the case of constructing an explanation, understanding results from a discursive game that the epistemic subject plays with herself[5]. In both cases, understanding is marked by a distinctive emotion – the feeling of satisfaction. This emotion, according to YP, plays an epistemic function. It signals the subject that

[4]YP further argue that this "seeing" encompasses two kinds of cognitive activities: reaching satisfaction by a successful analysis and by an insight.

[5]In seeking a new explanation, a person offers potential explanations to herself, evaluates them, discards, corrects and finally accepts one, again when she sees that an explanation is a correct solution.

she has gathered sufficient explanatory information, that this information is arranged in a satisfactory way, and that no further epistemic effort is needed.

On the other hand, by simply relating understanding to knowledge, Khalifa makes no epistemic distinction between believing[6] an explanation and (the reinforcing effect of) grasping that an explanation is a solution to an explanatory problem.

YP argue that this difference is essential, since it also signifies different epistemic investments. In knowing, one easily revises her beliefs in face of defeating information, whilst in understanding, the agent tends to "root for" or "insist on" the explanation, because of her previous investment in grasping it.[7]

The idea of differing in investment in knowing and understanding shall be important, firstly for the way in which YP characterize misunderstanding (see section 4), and the way in which I shall suggest both theories can be combined (see section 5).

2.2. The objective criteria for understanding

For Khalifa (2017), an explanation can be objectively correct, if it satisfies two clusters of criteria – global and local.

The *global* ones are that:

(1) it should be approximately true that q explains why p.

(2) q should make a difference to p.

(3) q should satisfy ontological requirements, of the explanation seeker.

The *local* constraints include: standards of the discipline, interests of the inquirer, and various requirements for structurally valid explanations given by theories of explanation. Khalifa then claims that local constraints should be satisfied in addition to global constraints.

For YP, an explanation is correct when it satisfies: *structural, informational and ontological requirements*. YP explicitly build their theory, based on Khalifa's claims,

[6]To the best of my knowledge, Khalifa (2017, 171) following Cohen (1992, 4) to define believing as: "belief that p is a disposition, when one is attending to issues raised, or items referred to, by the proposition that p , normally to feel it true that p and false that not-p, whether or not one is willing to act, speak, or reason accordingly." As opposed to this, grasping that *q explains p*, as supported by a specific phenomenal hue of satisfaction, disposes the epistemic subject to *act, speak, or reason accordingly*. For YP, *q explains p* is assimilated as an explanation when the subject *reasons* that *q explains p*, and this predisposes the person to *further reason* by the explanatory pattern of the explanation in similar cases (Yu & Petkov, 2024; Petkov, 2015). In a nutshell, believing can be epistemically ambivalent, whilst understanding as an outcome of a problem-solving activity is not.

[7]Recent studies in problem solving (such as writing an essay) using a search engine, LLM and brain-only support a similar picture (Kosmyna et al., 2025). Participants in the study that used LLM were much less inclined to critically evaluate the LLM's output, whilst participants who were in the brain-only group reported higher satisfaction and demonstrated higher brain connectivity, compared to other groups. One can easily argue that in both cases, subjects believe the output (that is, they have relevant concept possession). However, in the brain-only case, their emotional and cognitive involvement is much stronger.

as to deal with cases of explanations based on falsified but still used theories and idealizations. An important caveat is that their theory emphasizes that these criteria are not independently satisfiable.

For both theories, the objective criteria also lead to the characterization of two phenomena related to understanding – proto understanding and understanding based on falsified yet still used science.

The former, for Khalifa, arises from explanations that fall short of being correct, since they do not satisfy one's relevant ontological requirements, but play an indirect role for later correct explanations. The latter concerns explanations that can still be asserted as satisfying relevant criteria for approximate truth and ontological commitments, however coached in some contextually specific practical boundaries.

For YP's theory, the judgments are similar except for one caveat, since their theory emphasizes that understanding is a consequence of a problem-solving activity, the ontological status of an explanation, its informational content and structure are not independent. These criteria for YP are interdependent, which means for them, *global* and *local* constraints mutually influence each other, and cannot be satisfied independently. Consequently, it is possible to claim that the same explanation in a different context and historical period can offer understanding, misunderstanding, or partial or proto understanding based on falsified but pragmatically useful theories. That is, the same explanation can be objectively correct, incorrect, or partial solution to a problem. This depends entirely on the contextual constraints that the problem sets.[8]

What then are the different theories of misunderstanding that stem from these definitions of understanding?

3. Misunderstanding

The definition of misunderstanding that YP offer is:

[8]Following Laudan (1977), Danto (1985), and Mantzavinos (2016), YP similarly embrace the idea that scientific and epistemic standards are not a-historical. That is, the standards for successful problem solving and explanations are not separate from those that set ontological ones. They are interdependent. This follows from Danto's agnosticism for the existence of historical facts, independent from any narration about them. Combined with Laudan's pessimistic meta-induction, this historical contextualism permits a sort of *somber deduction argument*. In a nutshell, the idea is that one can first attest that one has no knowledge of the future. Consequently, one does not know how ontological and practical standards of future science would be. However, one *knows what current standards are* and has an abundant historiographical record of past science. Therefore, two judgments are available for past science: one based on the historiographical record, and one based on the current conceptual frameworks for scientific standards. These are both legitimate readings of past science, however they essentially resolve different explanatory problems. One is the problem of: "What have the standards of explanations and understanding been?" The other: "What are the current standards of explanations and understanding and from their perspective, what does past science look like?" The latter permits a potential *somber deduction* over the explanatory information about the same phenomena from past science, given the present standards, and offers a clear judgment of how close past explanations are from the current ones. This is a place where both realism and anti-realism can agree on.

> Misunderstanding is the epistemic failure, that occurs when a correct explanation is offered, that explanation can be in principle grasped (as the epistemic abilities and skills of the explanation seeker are sufficient for its grasping), but understanding of the explanandum through that explanation is blocked by a mismatch between the features of the explanation and the epistemic inclinations of the explanation seeker (Yu & Petkov, 2024, p. 20).

RK explicitly compare their claim with that of YP. For them misunderstanding is:

> Yu and Petkov claim that grasping incorrect explanations results in a lack of understanding, while we argue that *believing an incorrect explanation results in misunderstanding* (Rice & Khalifa, 2025, p. 7).

Another key difference is that RK focus not on what misunderstanding is, but on the corrective process that can make misunderstanding valuable – the processes which transform incorrect explanations to correct ones. They also add that, believing an incorrect explanation and misunderstanding can be valuable if:

> *S*1's misunderstanding possesses some epistemic value if at least one other agent *S*2's understanding is a reliably formed corrective response to *S*1's misunderstanding (Rice & Khalifa, 2025, p. 6).

Both theories then have a mismatch in their definitions of misunderstanding and in their aims. YP explore misunderstanding as rejecting a correct explanation, as well as what makes this possible. RK focus on characterizing valuable misunderstanding.

However, I believe that both theories can be fruitfully combined (see section 5). As we shall see, I argue that YP's theory defines misunderstanding more plausibly, whilst RK's idea of studying the corrective processes that follow misunderstanding can still hold.

3.1. YP's theory of misunderstanding

YP begin their analysis with the idea that a complete theory of understanding should cover a whole spectrum of related phenomena. The ordering which YP seem to offer to this spectrum is that, at the negative end lies *lack of understanding*, followed by explanations that clearly miss the mark, and then *misunderstanding*. On the positive side, they grade proto understanding, incomplete explanatory solutions, and finally proper understanding based on correct/complete solutions of explanatory problems. To clarify this, let's consider the problem: "Why did Merry dye her hair black?" (occurring in a pragmatic and psychological context) and the following explanatory answers:

A. We do not know.

B. Because aliens from planet X beamed her to do so.

C. Because she uses hair dye.

D. Because she bought a black hair dye.

E. Because she got motivated to embrace a gothic style after listening to Type O Negative's Black No. 1, bought and used the aforementioned dye color.

Note that A and all the possible answers of type B (all planetary systems where aliens could possibly exist and possess relevant technology), would not get us close to the correct explanatory answer. Both answers thus count as incorrect and do not extend our understanding. This is because saying nothing and saying the wrong thing equally do not advance us towards the right answer. C gets us closer, by being on the right track; however, it is still wrong, as it merely states the preposition of the explanandum. Nevertheless, we can perhaps count it as proto understanding, since it suggests the right context in which we should seek the answer. D counts as a partial solution, because it gets closer to the direct and complete answer E, which then offers the only correct explanation.

Why would YP put A and B in the same category of being incorrect, and also leading to a lack of understanding? It is in great contrast to claiming that B leads to misunderstanding, and that only A signifies lack of understanding, as RK's theory suggests.

On close scrutiny, A and B show that there are potentially an infinite number of incorrect but possibly valid answers to any factual explanatory question, and that this set includes epistemic silence. Note also that a logically invalid proposition, or one that does not form a proper explanatory inference with the explanandum, results automatically in an incorrect explanation. The set of invalid propositions is infinite, therefore there are an infinite number of incorrect explanatory answers to any well-formed explanatory question. Once we add the set of valid propositions (well-formed answers) that result in a false but possible answer, the number of incorrect explanations vastly outnumbers that of correct or partial ones. Consequently, we can argue that the epistemic outcome of incorrect explanations is exactly the same as the one for having no explanation whatsoever.[9]

To conclude, note that YP's theory offers a neat organization of the spectrum of understanding. This permits their theory to remain consistent with previously established categories – those of proto understanding and understanding based on falsified yet still used science. As a consequence, YP do not relate misunderstanding with explanations that fail to satisfy some of the objective requirements for being correct explanatory answer. Instead, they reserve misunderstanding for the failure that occurs when there is a mismatch between the objective criteria for a solution and the subjective expectancy for what would count as a solution. That is, when a potentially correct explanation is offered and an agent fails to grasp it, not because of lack of relevant knowledge or abilities, nor because the explanation is not potentially a correct answer to the explanatory problem, but because the said explanation does not satisfy her epistemic inclinations of what a correct answer ought to be.

These inclinations, as YP describe, are a form of bias. It occurs due to the emotional hue of understanding and tendency to seek a solution that conforms to previous

[9] I will return to this problem again when I discuss RK's idea that asserting incorrect explanations can be valuable (see section 4.2)

epistemic successes. Inclinations, YP argue, might block people from asserting an otherwise correct explanation, because it does not conform to such biases. This for YP can happen even when the person has relevant abilities, skills and knowledge to grasp the correct solution.

Importantly for YP's theory, misunderstanding can occur due to asserting both an incorrect explanation or a partial explanation, since inclinations from such explanations can block us from asserting a correct explanation. Hence YP's theory generalizes over all these cases.

For them then, misunderstanding firstly is a *subjective epistemic error* and secondly describes a specific contextual situation. That is a situation where the information for grasping a correct explanation is *available* or *within the epistemic grasp* of a person.

To summarize, YP's theory's main points are:

(1) Misunderstanding occurs only in the presence of skills and abilities to reach a correct explanation.

(2) Misunderstanding generalizes over asserting incorrect, partial or quasi explanations, when the relevant epistemic inclinations block agents from reaching a correct explanation.

(3) Misunderstanding is not a result of asserting an incorrect explanation alone. Neither having no-information nor having the wrong-information helps us to be on the right track towards a correct solution, hence both can be described as lack of understanding.

3.2. RK's theory of misunderstanding

RK do not focus on analyzing misunderstanding alone. More importantly, they wish to convey the idea that misunderstanding can be valuable. According to them, misunderstanding can trigger a corrective process which transforms one inquirer's misunderstanding to another inquirer's understanding. Such misunderstanding then becomes valuable for the inquirers that correct it. Nevertheless, to describe these corrective processes, RK must first draw the line between misunderstanding and lack of understanding. For them, lack of understanding concerns cases where one has no explanatory information whatsoever, whilst misunderstanding occurs when one believes in an incorrect explanation. This is one key difference between the two theories of misunderstanding.

RK also claim that, neither does a misunderstander need to abandon her mistaken explanation, nor does the knowledge required for a correction need to be available to her. A mistaken explanation can be amended by others at any later time and at no cost to the epistemic subject that has propagated the mistake (Rice & Khalifa, 2025). This is another key difference between the two theories.

According to RK their generalization of misunderstanding as believing the wrong explanation and describing relevant corrections as available at no cost (of the misunderstander) and at any later time, makes their theory more generally applicable than that of YP. They also suggest that it might also discriminate more plausibly between misunderstanding and lack of understanding. Nevertheless, I believe that in light of the distinctions that YP make this move is not justified (see section 3.1 and 4.2).

Crucially, RK do not make a distinction between valuable and non-valuable misunderstanding, and thus between different types of mistaken explanations. They also do not offer necessary and sufficient conditions for valuable misunderstanding, by claiming that there could be many corrective processes and their preliminary goal is to describe only three of them. They only provide sufficiency conditions for misunderstanding. The three processes, by which people can reach understanding when correcting mistaken explanations, require them to:

(1) Consider mistaken explanations only as potential explanations.

(2) Draw true counterfactual claims from investigating mistaken explanations.

(3) Develop methods and discover evidence that permit them to distinguish mistaken (yet possible) explanations from correct (hence actual) explanations.

RK illustrate these three processes by three historical cases of mistaken explanations: intelligent design as an explanation of evolution, explanations of autism as caused by vaccines, and the dispute between Darwin and Kelvin. For brevity, I discuss only two of them: Darwin vs Kelvin (Burchfield, 1990) and autism caused by vaccines (Wakefield et al., 1998; Rao, 2011).

RK do not offer explicit explanatory problems as a framework to reconstruct the dispute between Darwin and Kelvin, but I believe it is useful to at least roughly try to frame the dispute as an explanatory one. Darwin explained species formation as a gradual process of phenotype change via natural selection (Darwin, 1859). This explanation entails that in order for the gradual process of selection to hold, the age of the Earth should support a very large number of biological lineages. Darwin never gave a specific estimate of the age of the Earth, nor did he offer any explanation for such an age. The age of the Earth was an element in his explanans and not an explanandum for him. That is, if evolution by natural selection is sketched as a causal explanation, then the age of the earth would be a supportive cause to the effect of the existence of multiple species.[10]

Kelvin, on the other hand, favored a theistic evolution, which permitted much faster species formation (Bowler, 1983). He was not a biologist but a mathematical physicist and an engineer. His scientific focus also wasn't on explaining species formation. Instead, his estimates of the age of the Earth were driven by an interest in thermodynamics and the shape of the planet. His views on the age of the planet changed over time, yet he ultimately settled on 20–40 million years. A key point in his calculations was that the Earth is uniform in composition and cools off as a rigid solid.

[10]Since typically effects do not explain causes, the presence of multiple species and the process of natural selection cannot explain the age of the earth.

From this brief outline we can note that, in one sense, Darwin and Kelvin had two different epistemic projects. Darwin explained evolution, and Kelvin made a hypothesis about the age of the planet and justified it using the physics of his time. In another sense, however, they had two competing explanations of evolution (natural selection vs. theism) and two inconsistent factual claims about the age of the Earth.

RK do not make these distinctions explicit. Rather, they focus only on the claim that Kelvin had a mistaken estimate of the age of the Earth. They further propose that his misunderstanding was valuable for both physics, geology and biology, as it pushed scientists to seek correct solutions. Physicists (Perry, 1895) showed that if the Earth is not rigid and homogeneous, its age could be much older (up to billions of years). Darwin himself began thinking about sexual selection as one process that could expedite evolution (Goodman, 2019).

According to RK, besides both seeking a better evolutionary theory and a more precise estimate of the age of the Earth, Kelvin's misunderstanding was valuable because scientists also gained counterfactual information: "How old the Earth would be if Kelvin's assumptions were correct".

RK draw a similar picture for their other case study – that of explaining autism by vaccination. In 1998 Wakefield and his colleagues performed a study, according to which there is a link between the measles, mumps, and rubella (MMR) vaccine and developmental disorder in children, i.e. autism (Wakefield et al., 1998). Their causal explanation was that these vaccines cause enterocolitis, which in turn leads to disorders in neurological development. Despite the small sample size (n=12), the uncontrolled design, and the speculative nature of the conclusions, the paper received wide publicity. Parents were concerned about the risk of autism, and MMRR vaccination rates began to drop. Eventually, Wakefield's study was retracted and criticized for its small sample size and poor methodology. Later studies showed that even as a hypothesis, this causal vector of autism was implausible (for a short overview and references, see Rao, 2011).

RK claim that Wakefield's explanation functioned as a valuable misunderstanding, because it permits two out of three corrective processes. One is to consider alternative explanations, to quite:

> Finally, the scientific community responded to Wakefield's claims by considering vaccines as a potential cause of autism and proposing several alternative hypotheses that might account for the results (e.g., poor sample size or spurious correlation) (Rice & Khalifa, 2025, p. 8).

And one is to gather counterfactual information:

> Finally, autism researchers considered three mechanisms whereby vaccines might cause autism: encephalopathic proteins caused by damaged intestinal lining; neurotoxicity of certain preservatives found in vaccines; and overwhelming the immune system... (Rice & Khalifa, 2025, p. 9)

Both claims immediately invite criticism, since RK's definition of valuable misunderstanding is not satisfied by Wakefield's research as a case study. As we saw above,

for RK misunderstanding is valuable, if some other agent's *understanding is a reliably formed corrective response* to an incorrect explanation. That is when one forms understanding (i.e. obtains a correct explanation) as a consequence of such corrective process. However, since no correct explanation is offered to "Why do vaccines cause autism?"[11], nor to "Why is there autism?" none of the corrective responses actually provide understanding – the kind of understanding described by Khalifa in his previous work, as knowledge of correct explanations (see section 2). But RK claim that in this case understanding is brought by "gaining counterfactual information", that is, by understanding the potential mechanisms through which vaccines *could* cause autism. However, as we saw (in section 3.1), there is a good argument for why incorrect explanations (even possibly true) do not extend our understanding. As such, I believe YP's theory can offer a much better assessment of Wakefield's explanation, (at best) as leading to lack of understanding or (at worst) as sidetracking potentially positive research on autism. This leads us to the next section on criticisms of both theories.

4. Criticisms of two theories of misunderstanding

In this section, I will present some criticisms of the two theories of misunderstanding. My criticisms mainly concern two questions:

(1) Are these theories internally coherent?

(2) Are these theories rich enough to capture all related phenomena?

The critical problem the first question poses is: whether these theories keep the existing distinctions intact, when they define misunderstanding and lack of understanding. These existing distinctions should include proto understanding, and understanding based on falsified yet still used science. The second question concerns whether these theories map the whole spectrum of understanding, which questions whether there are related phenomena that remain outside of the scope of any of the theories.

My short answer is that RK fail to properly address the first problem, which then presents a significant challenge to their theory as a theory of misunderstanding. Nevertheless by focusing on corrective processes their theory offers a broader scope. On the other hand, YP's theory can properly address the first problem but not the second, in that the theory is internally coherent but does not properly address the dynamics of the corrective processes that stem from misunderstanding.

4.1. Shortcomings of YP's theory

RK point out that YP's theory has two main problems: a lack of generality, and rigidity. Here I agree with the first problem, but not the second. I shall argue that YP's characterization of misunderstanding is more plausible than that of RK. As such, YP's theory does not suffer from being too rigid.

[11] No correct factual explanation is possible given the false preposition of the why question (see van Fraassen, 1980, pp. 139–140 for similar treatment of what he calls foolish and dumb questions). For further analysis of the question in a modal context, related to our world, see section 4.2.

RK believe that their theory more plausibly distinguishes misunderstanding from lack of understanding. To quote:

> Yu and Petkov (2024, 64) claim that grasping incorrect explanations results in a lack of understanding, while we argue that believing an incorrect explanation results in misunderstanding. In many cases, one lacks understanding because one has no beliefs about an explanation of a phenomenon. For example, many middle-aged men feign no hypotheses about Taylor Swift's latest fashion choices. It seems more plausible to say that they lack understanding of Swift's wardrobe preferences than that they misunderstand these preferences (Rice & Khalifa, 2025, p. 7).

This critique is incorrectly placed. As we saw (in section 3.1), for cases where no explanation is offered and sought, YP's theory would claim that the person lacks understanding, not that she misunderstands.[12] This is the same epistemic outcome as described by RK.

A further critique from RK is that YP's theory is overly restrictive, since RK claim that individuals can misunderstand even if no correct explanation is available. They illustrate this with the example that both Kelvin and Wakefield misunderstood, due to them believing incorrect explanations, even when no correct explanations of their respective explananda had been offered.

This critique is also not well placed, as we shall see for Kelvin's case there was a better explanation available that Kelvin rejected – that of his assistant Perry. Moreover one can also argue that Kelvin's understanding of the age of the Earth could be seen as a form of proto understanding (see section 4.2) and not of misunderstanding. Both these perspectives are well within YP's theory's scope.

Considering Wakefield's explanation, YP's theory would render it as a lack of understanding, since no correct explanation is possible for a false explanandum (see section 4.2). I shall further argue that Wakefield's case constitutes a significant problem for RK. This is because the correction of Wakefield's case essentially consists of rejecting the explanandum altogether and not of arriving at any correct explanation and thus of understanding of autism.

Finally as we shall also see (section 4.2) Laudan's pessimistic meta-induction poses a significant challenge to the claim that an explanation can be incorrect and lead to misunderstanding in an ahistorical context, that is when a correct explanation might become available in.

Given these points, I believe YP's theory lacks generality not because it fails to properly distinguish between various epistemic outcomes, but because it is not sufficiently descriptive. That is to say, YP, as opposed to RK, do not make a distinction between different types of corrections. As such, RK's theory more plausibly describes different corrective processes.

[12] When offered a correct explanation, an agent who has no previous relevant beliefs would also lack an inclination to reject it.

But for now, we can conclude by asserting that RK's critique has some merit. YP's theory is indeed too narrow. Their theory does not acknowledge the different consequences of having the wrong explanation. That is grasping the wrong explanation could both lead to misunderstanding as rejecting a correct explanation but also not to seek a correct explanation to begin with. As we saw (in section 2.1), to YP the positive phenomenology of grasping signals to the person to stop any further inquiry into the problem. Consequently, a person that grasps an incorrect explanation might not attempt to discover the correct explanation, because she already has accepted an explanation that is coherent with her current beliefs. RK's approach is much better equipped to handle such cases. For them, the corrective process is the focus of misunderstanding. This suggests that one's healthy skepticism could play a positive part in eventually reaching positive epistemic outcomes (see section 5).

4.2. Shortcomings of RK's theory

I see two serious flaws in RK's theory. One involves the problem that, as it stands, their definition of valuable misunderstanding can lead to a mismatch of judgments on whether some incorrect explanations generate misunderstanding, proto understanding, or a complete lack of understanding. The other main criticism is that the corrective processes they suggest are too loose. It is easy to illustrate both problems with RK's own examples: Kelvin VS Darwin and Wakefield's explanation of autism.

Considering Kelvin's hypothesis about the age of the Earth and its explanation, we can also frame his results (taken separately from the critique of Perry) as generating proto or quasi understanding. To see this, we must first note that RK claim that the difference between ¡idealized and therefore strictly speaking false explanations¿ and ¡incorrect explanations¿ is that incorrect explanations involve some false central beliefs about why the phenomenon has occurred (Rice & Khalifa, 2025, 6). Looking closely at Kelvin's explanation as to why the Earth is 20-40 millions years old, we have two central assumptions. One is Fourier's law, and the other is that the Earth is a rigid solid. Fourier's law is still in circulation, whilst the hypothesis about Earth's composition is factually incorrect. Given the best scientific knowledge of his time, one of the two central assumptions for his hypothesis could be taken as being correct. We also saw (in section 2.2) that Khalifa defines proto-understanding as a result of explanations that are on the right track. By using Fourier's law, Kelvin's hypothesis can be taken to be on the right track towards better and more accurate ones. Thus, one can judge that Kelvin actually had proto understanding of the age of the planet and did not misunderstand it, since the laws of thermodynamics are essential for understanding its age and cooling.

This critique can be extended even further. Recall that RK's theory of misunderstanding defines explanations as incorrect, when any future explanation falsifies some of their central assumptions. This claim makes a theory of understanding vulnerable to Laudan's pessimistic meta-induction, because any future science is likely to reject the central assumptions of past explanations.

Returning to the debate about Kelvin's estimates about the age of the Earth, I believe YP's theory offers a much better framing. Their theory also suggests that Kelvin misunderstood the age of the Earth. That is, he had a partial or quasi explanation for the age of the planet, but rejected a better hypothesis that was within his epistemic grasp. To clarify:

Kelvin correctly pointed out that Fourier's law is relevant, but he could not have known relevant facts about radiation which would have certainly changed his estimates. On that account, it makes more sense to say that he lacked relevant knowledge of radiation, and that's one reason for his wrong estimates, whilst by pointing to the relevance of Fourier's law, he was on the right track.

However, looking closer at the science available to him at the time, some have also suggested that his thinking was fundamentally flawed, because he rejected the estimates of his assistant Perry (Shipley, 2001; England et al., 2007). Here investigating the debate between Perry and Kelvin from the perspective of YP's framework renders a cleaner historiographical reading. Firstly, because it involves a dispute within physics and the emerging geology, and not between two (at the time) disconnected disciplines, i.e. physics and evolutionary biology. Secondly, because the historiographical record of publications at the time clearly shows that Perry tried to improve on Kelvin's explanation (Perry, 1985) and that Kelvin considered and rejected Perry's estimates (Kelvin, 1895). As such, this debate presents a clear case of rejecting a correct (given the physics of the time) explanation that was within the epistemic grasp of Kelvin. Therefore, considering the problem of Kelvin's understanding of the age of the Earth, it seems to me that it is better to describe his lack of knowledge of radiation as ¡lack of understanding¿, and his missed opportunity to consider the Earth's specific structure as ¡misunderstanding¿, and his consideration of Fourier's law as granting him ¡proto or partial understanding¿. Given this reading, Darwin's arguments for much older Earth would have added only further weight for doubt in Kelvin's initial estimates. However, again Kelvin's commitment to theism can be seen as an inclination that blocked him from asserting Darwin's evolution by natural selection and thus considering that the Earth could be much older.

As such, I believe YP's theory provides a much cleaner and more accurate historiographical reading of the debate, as it gives a crisp order of the spectrum of understanding, lack of understanding and misunderstanding.

The second major problem for RK's theory is that, their judgments of what makes misunderstanding epistemically valuable hinges on corrective processes that are too permissive. Essentially all the corrective processes they list involve asserting incorrect explanations as possible. This permits defining valuable misunderstanding simply as knowledge of possibility.

However, as RK themselves note, inferring true possibilities from false explanations is straightforward (Rice & Khalifa, 2025, p. 11). The only thing one needs to do,

if an explanation is logically valid,[13] is to strap a diamond to it and the whole inferential structure becomes true and correct (in a possible world). However, a corrective process that simply points out that an explanation is possibly true, by denying that it is actually true, achieves very little epistemically.

This process consists of simply rejecting at least one proposition of the explanation by showing that it is inconsistent with a fact. This can happen without actually extending our knowledge of the facts that lie at the center of the explanandum. This is especially problematic if the explanandum can be reconstructed as a conjunction between a fact and a fiction. It is easy to illustrate this problem with Wakefield's case. The problem "Why vaccines cause autism" can be broken into the factual: "There is autism in infants" and the hypothetical "Vaccines cause autism in infants".

Since the explanandum is factually false, any valid explanans would result in a possibly true structure. We can have any number of modally valid inferences "*because vaccines have neurotoxic properties X*", all of which would result in incorrect explanations.[14] Considering one such property, then rejecting it as factually incorrect, then moving to another possibility would not necessarily lead us an inch closer to actually explaining autism.

In fact this dynamic is exactly what we see in the studies that followed Wakefield's. They rejected the validity of several hypotheses "Why vaccines cause autism", and did that without providing any novel explanation to the fact "Why there is autism". No correct explanation (of the cause of autism) was ever offered, since no correct explanation in this context is possible to begin with. Instead, researchers simply rejected the explanandum altogether.

Nevertheless, RK paint Wakefield's case as offering valuable misunderstanding, because according to RK researchers have gained counterfactual information. Yet this violates Khalifa's own definition of explanatory understanding, as knowledge of a correct explanation, since no correct explanation is possible when the explanandum is false. To clarify, a corrective answer to a false explanatory preposition, that consists of an actual explanation, must go beyond the possibility space of the incorrect 'why' question. Such a correction has to first reject the explanandum, then paraphrase it in a possibility space where there could be an actual answer, and finally arrive at this answer. Such a correction however is not strictly necessary for a false 'why' question (see van Fraassen, 1980, pp. 139–140).

Not only that, the claim of RK also leads to a picture of scientific progress very much like the Red Queen's race. To have scientific progress, one simply needs to relate a fact to a fiction in an explanandum and provide a valid fictional explanans. Since a fictional explanandum modally forks to an infinite number of equally plausible explanations, i.e. all the possible worlds in which there is a causal vector between autism and vaccinations, exploring this infinite set of possible worlds would count

[13] In fact, friends of paraconsistency can say that an invalid explanation can be true in the world where logic fails.

[14] These explanations would be incorrect, since they violate relevant ontological commitments when they are reasonable, as Khalifa (2023) claims.

as "gaining counterfactual information", and make each correction valuable. This is because in each rejection of such an ¡explanandum/explanans¿ pair a correction would need only flag some of the propositions of the explanans as false, thus not necessarily exposing the culprit – the fact that the explanandum itself is wrongly postulated.

Instead, it is more plausible to propose that a corrective process should offer a correct explanation in place of a false one. That is an explanation that satisfies structural, informational and ontological requirements, and thus elucidates facts in our world.

To summarize, the problem with the corrective processes, as described by RK, is that they fail to guarantee correct explanations under Khalifa's own definition. This is because flagging possibilities as factually false does not necessitate a factually correct explanation as a form of correction.

Nevertheless, I believe RK's idea that valuable misunderstanding can be easily amended. One can, for instance, distinguish between two types of corrective process, which are fundamentally different:

(1) The processes that flag as false the preposition of the explanandum.

(2) The processes that flag as false some of the propositions of the explanans.

By now it should be easy to see that the latter would be epistemically more economical and could lead to arriving at better explanations.

To conclude, although I believe RK's account presents an interesting approach to analyzing the processes that correct misunderstanding, their theory suffers from being too loose. On the one hand, it blurs the existing categories of proto-understanding and understanding based on falsified theories. On the other, it too loosely permits any correction to render the wrong explanation, as epistemically valuable, as a correction (and thus reach understanding).

5. Clearing misunderstanding towards a unified theory

From the criticisms raised in the previous sections, it should be clear that both theories are partial. Hence, a unified picture that combines the positives from both can be obtained. On one side, we can consider that YP's theory characterizes misunderstanding better. On the other, RK's theory can be seen as offering a valuable insight into the corrective processes that transform misunderstanding to understanding. If we assert that misunderstanding is a rejection of a correct explanation, we can easily have the following addendum:

> Misunderstanding between two parties *is resolved via corrective processes that establish a correct explanation* in place of a partial or incorrect one.

What would such a process entail? Returning to our framing of understanding as having *objective* and *subjective* dimensions can be useful. On the objective side, a corrective process needs to establish a correct explanation in place of an incorrect or

partial one. Assuming that the incorrect explanation has a valid explanatory structure[15], we saw that there could be only two options where such an explanation could be wrong:

(1) The explanandum might be false.[16]

(2) The explanans could contain some or only false propositions.

In the first scenario, offering a correct explanation would reject the explanandum by framing the problem properly. For instance, in Wakefield's case, the correction is constructed as a negation of the explanandum itself, as factually implausible, and rejects his explanans as methodologically flawed. Note also that if no explanation is offered to the central explanatory problem, "Why there is autism?" then the correction simply has shown that Wakfield's explanation was wrong. That is, it did not provide any understanding, yet also since no novel explanations have been offered, our understanding of autism did not increase.

In principle, however, a corrective answer can both re-frame the explanandum properly and provide a correct explanation. That is, a future correct explanation of autism would explain what caused it in the subjects of Wakefield's research. Yet that explanation would not include vaccines, nor any of the vectors of autism that Wakefield's supporters have asserted. As such, it is safe to claim that Wakefield's case would be more properly framed as explanations that lead to a lack of understanding (of the actual causes of autism).

In the second scenario, a corrective explanation would include asserting as true[17] relevant propositions in the rejected explanans (if any), and replace only the incorrect ones. Kelvin's case presents an illustration of that process. That is, a correct estimate of the age of the Earth includes laws from thermodynamics that Kelvin himself considered, and rejects the idea that the Earth is a uniform solid.

On the subjective side of clearing misunderstanding, if we assert that misunderstanding concerns rejecting a correct explanation, we have two sides in a debate:

(1) For the subjects who establish the correct explanation, the misunderstanding can be valuable.

(2) For the subjects who reject the correct explanation, clearing their misunderstanding requires overcoming relevant epistemic inclinations.

The first side is what RK's theory characterizes. It is obvious that both framing an interesting explanatory problem and offering a partially correct explanation can be valuable. The former can spur novel research, if the scientific community asserts that the problem is worth exploring. The latter (when a partial or proto-explanation is offered) can be seen as being on the right track towards a correct explanation. Both situations are well within the scope of Khalifa's theory of understanding and

[15]Valid here means that the explanation obeys structural and informational requirements for a candidate solution to the explanatory problem.

[16]i.e. would fail under some reasonable ontological requirements, if one wishes to weaken the more plausible claim that a scientific explanandum should be factual.

[17]Here partial, pragmatic or approximately true theories can potentially serve as an evaluation, if one does not embrace an outright factivity for explanations.

RK's characterization of valuable misunderstanding, if the latter characterization of corrective processes is made tighter.

We saw that there are significant logical and epistemic arguments, which deter us from asserting that just considering wrong explanations as offering valuable counterfactual information is sufficient for understanding. Instead, it is more epistemically meaningful to assert that corrective processes should result in establishing the (actually) correct explanation in the place of an incorrect (but possibly true) one.

This leads us to the final side of clearing misunderstanding, when a subject needs to clear her misunderstanding by asserting a correct explanation which she has initially rejected. Here, I should emphasize that asserting a correct explanation from testimony and direct acquaintance are both realistic epistemic possibilities. In the former, it is often the case that the explainer is tasked with not only conveying the correct explanation, but also pointing out where the incorrect one has gone astray. The misunderstander, then, needs to shed light on her own epistemic inclinations and be ready to make necessary belief revisions. Here, openness as well as the ability to accept criticism and critically examine personal epistemic stances would be valuable. On the other hand, mental rigidity as well as the inability to investigate other's arguments are detrimental for clearing misunderstanding.

Finally, it is also possible for an epistemic agent to make such epistemic revisions by herself.[18] Misunderstanding can also be framed more minimally as the situation where the path to reaching a correct explanation is blocked, even though abilities and skills for reaching such an explanation are present. As we saw, misunderstanding occurs as a consequence of the positive reinforcing role of the sense of understanding (Yu & Petkov, 2024). It is then possible (as the critics of the sense of understanding point out (Trout, 2002) to be misled into asserting an incorrect or partial explanation simply by the assuring feeling of insight that accompanies such an explanation. This can happen when, at a later time, we might assert that this explanation was incorrect, and arrive at a correct one. That is in a situation where we, as epistemic agents, have the necessary skills and abilities to solve the explanatory problem by ourselves. The following epistemic dynamic familiar probably to all philosophers can serve as an illustration. Waking up early in the morning, we feel excited for a new solution to a vexing problem with which we have struggled. But by the time we finish our morning coffee, we consent that the solution in which we were previously so sure would inscribe our names in the pantheon of philosophy, is actually a no go. Yet at later time, when fully focused, we could arrive at an insight to that problem and a solution that is truly valuable.

What is the cognitive underpinning of such a false sense of insight? Firstly we should note that the literature on insight solutions is currently converging on the idea that the sense of insight is reliable yet fallible. For instance, Laukkonen et al. (2020) suggest that insight solutions depend on a coherence between our beliefs and the vexing situation (an explanatory problem in our case). As such, a false sense of

[18]Special thanks to Krasimira Filcheva and Marina Bakalova for the insightful discussion on the possibility of such an epistemic situation.

insight can be experimentally produced by semantic priming (Laukkonen et al., 2020). Relevant to our topic, this empirical result shows that if an agent has false beliefs or if her beliefs are partial to the information required to solve an explanatory problem, she can arrive at a solution that is partial or incorrect and have a misleading sense of insight. Note also that the sense of insight as a mechanism would not be at fault here, since it has not malfunctioned. That is, it has correctly pointed us to a solution that increases informational coherence, yet the information it operated upon was false to begin with.

What is important for us, then, is not the problem if the sense of insight is or is not reliable, but instead that a false sense of insight can deter us from searching for a correct solution (at least temporarily). Only after we consider the solution again, this time focusing our mind on a larger set of relevant beliefs, can we discover that the solution we had was no longer correct (i.e. coherent) and make space for a more accurate insight to occur, as we arrive at a correct explanation.

To conclude, I believe that a coherent theory of misunderstanding and its related corrective processes, is a step in the right direction towards a complete theory of understanding.

References

Baumberger, C. (2019). Explicating objectual understanding: Taking degrees seriously. *Journal for General Philosophy of Science*, 50(3): 367–388. doi:10.1007/s10838-019-09474-6

Bowler, Peter J. (1983). *The eclipse of Darwinism: anti-Darwinian evolution theories in the decades around 1900*. Baltimore: Johns Hopkins University Press.

Burchfield, J. D. (1990). *Lord Kelvin and the Age of the Earth*. University of Chicago Press.

Cohen, L. Jonathan. (1992). *An essay on belief and acceptance*. Oxford: Clarendon Press.

Danto, A.C. (1985). *Narration and Knowledge: (Including the Integral Texts of Analytic Philosophy of History)*. New York: Columbia University Press.

Darwin, C. (1859). *On the origin of species*. John Murray.

England, P. C., P. Molnar and F. M. Richter. 2007. John Perry's neglected critique of Kelvin's age for the Earth: A missed opportunity in geodynamics. *GSA Today* 17(1): 4–9.

Goodman, L. E. (2019). *Darwin's Heresy Philosophy* 94 (1): 43–86.

Kelvin, L. (1895). The age of the Earth. *Nature*, 51(1323): 438–440.

Khalifa, K. (2017). *Understanding, explanation, and scientific knowledge*. Cambridge University Press. doi:10.1017/9781108164276

Khalifa, K. (2023). Should friends and frenemies of understanding be friends? Discussing de Regt. In I. Lawler, K. Khalifa, & E. Shech (Eds.), *Scientific understanding and representation: Modeling in the physical sciences*. Routledge.

Kosmyna N, Hauptmann E, Tuan YT, et al. (2025). Your brain on chatgpt: accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv* doi:10.48550/arXiv.2506.08872

Laudan, L. (1977). *Progress and its problems: Toward a theory of scientific growth*. Berkeley, CA: University of California Press.

Laukkonen, R. E., Kaveladze, B. T., Tangen, J. M., & Schooler, J. W. (2020). The dark side of Eureka: Artificially induced Aha moments make facts feel true. *Cognition*, 196, p. 104122.

Lawler, I., Khalifa, K., & Shech, E. (2023). *Scientific understanding and representation: Modeling in the physical sciences*. Routledge.

Mantzavinos, C. (2016). *Explanatory pluralism*. Cambridge, England: Cambridge University Press.

Petkov, S. (2015). Explanatory unification and conceptualization. *Synthese*, 192(11): 3695–3717. doi:10.1007/s11229-015-0716-2

Petkov, S., & Yu, H. (2023). What is scientific misunderstanding? *Journal of Human Cognition*, 7(2): 5–18. doi:10.47297/wspjhcWSP2515-469901.20230702

Perry, J. (1895). On the age of the Earth. *Nature*, 51(1314): 224–227.

Rao TS, Andrade C. (2011). The MMR vaccine and autism: Sensation, refutation, retraction, and fraud. *Indian J Psychiatry*. Apr; 53(2): 95-6. Doi:10.4103/0019-5545.82529.

Rice, C., & Khalifa, K. (2025). Thank you for misunderstanding! *Philosophical Studies*. doi:10.1007/s11098-025-02311-1

Sharlin, H. I. (1979). *Lord Kelvin: The Dynamic Victorian*. Pennsylvania State University Press.

Shipley, B. C. (2001). 'Had Lord Kelvin a right?': John Perry, natural selection and the age of the Earth, 1894–1895. In C. L. E. Lewis & S. J. Knell (Eds.), *The age of the Earth: From 4004 BC to AD 2002*. London, England: Geological Society. doi:10.1144/GSL.SP.2001.190.01.08

Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, 69(2), pp. 212–233. https://doi.org/10.1086/341050

van Fraassen, B. C. (1980). *The Scientific image*. Oxford University Press.

Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., Berelowitz, M., Dhillon, A. P., Thomson, M. A., Harvey, P., Valentine, A., Davies, S. E., & Walker-Smith, J. A. (1998). Retracted: Ileal-Lymphoid-Nodular hyperplasia, Non-Specific colitis, and pervasive developmental disorder in children. *The Lancet*, 351(9103): 637–641.

Yu, H., & Petkov, S. (2024). Don't get it wrong! On Understanding and its negative phenomena. *Synthese*, 203(2), 48. doi:10.1007/s11229-023-04445-3

Balkan Journal of Philosophy (2026), vol. 18, issue 1, 86 – 98

WHY ARE PRACTITIONERS NOT REALISTS ABOUT ANIMAL MODELS OF ANXIETY AND DEPRESSION?

Andrei Ionuț Mărășoiu*

Abstract: Researchers and practitioners rarely take on a realist attitude about animal models in psychiatry. Why does this happen? I will argue that even the best among current models fail to satisfy two requirements that demand care in balancing them one against the other: the psychological construct requirement and the mechanistic requirement. Failure to satisfy them explains why researchers do not believe in them. I conclude by arguing that, if one were to be a realist about animal models in psychiatry, interventionist and representational aspects of that attitude would be inextricably mixed.

Keywords: animal models, psychiatric conditions, analogical reasoning, scientific realism, model validity.

1. Introduction

Researchers and practitioners rarely take on a realist attitude about animal models in psychiatry. Why does this happen? The question is of immediate concern to psychiatrists and experimental neuroscientists. It also clearly targets the nature of the mind, as it is gleaned from models of psychiatric disorders. And the details and limitations of animal models speak to the distinctive role animal models play in psychiatry, as opposed to more theoretical models in other fields of scientific inquiry.

Why not be a realist about current animal models in psychiatry? I will argue that even the best among current models fail to satisfy a set of demanding requirements, and such failures explain why researchers do not believe in them. But this explanation leaves the door open for some future animal models that may satisfy such requirements. I conclude by arguing that, if one were to be a realist about animal models in psychiatry, interventionist and representational aspects of that attitude would be inextricably mixed.

By an animal model I will understand the successful/corroborating testing on animals (e.g., rodents) of conjectures meant to apply to human beings. Animal models have gained wide currency in branches of medicine as different as psychiatry and dental health (La Follette and Shanks, 1995, pp. 142-143). The relation between animal models and their human targets has been described in different ways. Animal models have been said to be designed based on *analogies* between animals and humans (La Follette and Shanks, 1995, p. 147), for the purpose of *simulating* human disease in the animals (Piotrowska, 2013, p. 440), *controlling* the evolution and parameters of the animal disease, and then *extrapolating* attempted cures from animals to humans

*University of Bucharest; andrei.marasoiu@filosofie.unibuc.ro

(Steel, 2010, p. 1058). These characterizations are not equivalent. Analogies between rodents and humans can also be the product of homology of structures, or of shared genes, as opposed to intervention-induced simulation of disease. And controlling parameters in animal models far outstrips what can be extrapolated from animals back to the human case.

Representation in animal models, if achieved, can be of several kinds (Piotrowska, 2013, p. 445). It can be *behavioral* (symptoms or impairments of animals stand for those of humans), it can be *structural* (structures causally responsible for behaviors in animals stand for those in humans), it can be *aetiological* (environmental or structural factors in animals carry over to the human case), or *therapeutic* (medication or behavioral therapy successful in animals might be possible for humans). An intuitive demand might be that therapy, structure and behavior should be causally correlated (La Follette and Shanks, 1995, p. 147), though only correlations (not causation) are available as experimental evidence.

Practical concerns are extremely important in the design and use of animal models. These applications include the possible cure for a disease, but also prevention by screening tests (Willner, 1984, p. 11). Researchers are partly motivated by seeking funding opportunities from private companies (Cartwright 1991, p. 65) or authorization to carry out experiments from the FDA and to move from the animal model stage to clinical trials (Chader, 2002, p. 393). These different uses to which animal models are put do not detract from the important question of representation: while what a model is good for and what that model represents are very different questions (Morrison, 1999, p. 46), a realistic stance may legitimately ask about the extent to which successful representation is a guide to successful application outside the laboratory (*contra* Cartwright, 1991, p. 66). If laboratory intervention occasions some hypotheses about the animal models used, what does it mean to say that an animal model represents its human target?

2. Construct validity

Assessment of animal models of disorders typically involves a threefold distinction between kinds of validity (Willner, 1984, p. 1). Predictive validity of a model consists in the ability to make predictions based on the animal model that carry over, to some degree, to the target human system. Face validity of a model consists in the ability to replicate human symptoms in the animal model and rescue the induced effects by behavioral training or medication. A model's construct validity consists in the ability to replicate not phenotypical traits, but their hypothesized neural or genetic structural causes.

Practitioners (e.g., Willner, 1984, pp. 1–2) sometimes act as though prediction, face, and construct validity are different dimensions along which to assess the degree of success of animal models. Yet animals are organisms, i.e., integrative systems (Piotrowska, 2013) where behavior, neural structure, and physiology are interrelated. This shows in the interaction between different kinds of validity. For example, we should expect low scores on both predictive and construct validity if face validity does not obtain, since if not even the phenotypical trait can be represented, then it is hard to

see how structures responsible for *it*, or medication meant to improve *its* impairments, can be provided.

Predictive validity makes claims of predictive success, and face validity makes claims of empirical adequacy (up to a degree) of the animal model to the target human system. Construct validity seems to make the claim that, if face validity obtains (up to a degree), then the structures responsible for those behavioral effects in animal and human are analogous. So, of the three kinds of validity, only construct validity makes a representational claim that goes beyond empirical adequacy, namely, that the neural and genetic manipulations in the animal model could be blueprints for designing analogous treatments in the human case. (Still, the interrelatedness of the three kinds of validity noted above makes the issue of representation relevant to all three.)

There are many measures of overall model validity. For instance, Willner (1984, pp. 1–2) proposes the strength of the correlation between test potency (on animals) and clinical potency (on human patients). More recent work (Strauss and Smith, 2009, pp. 16-19) suggests an increasing complexity from statistical and methodological standpoints. The different scores on construct validity will have the consequence that one can only speak of models that capture theoretical constructs *more or less*, rather than fully or not at all.

Issues of measurement aside, construct validity is a notion borrowed from general psychology and later psychometrics. As Strauss and Smith recount (2009, p. 5), the validity of a construct was initially the validity of a *psychological* construct. Examples of psychological constructs include disorders like anxiety or depression, but also more general constructs like fear or satisfaction, or more behaviorally and physiologically testable ones, like anger or sleeplessness. Naturally, the notion of construct validity is generalizable so that the construct could stop being a psychological one. But this is not possible in psychiatric models, where what is at stake is precisely how to model in animals a human psychological phenomenon. The reason for calling it a construct comes from the desire to distinguish unique clinical phenomena from their *taxa* belonging to disease classification and operative in deciding on a diagnostic. So, any model of a psychiatric disorder has to worry about the validity of how it models some core psychological construct. Animal models make no exception.

3. Psychological reality in animal models

I will focus on an example. In Choi et al. (2012, pp. 1–3), fear is induced and consolidated behaviorally in mice by associating tones heard by the mouse with foot shocks. How well fear will have been induced and consolidated is measured by how often the mouse freezes (or decreases locomotion relative to baseline) when hearing tones not accompanied by shocks. As it turns out, by a molecular manipulation, namely, decreasing the level of BDNF (brain-derived neurotrophic factor) in their prelimbic cortex (by gene knock-down or by a targeted viral infection), mice have a diminished ability to consolidate their fear memories (they behave fearfully for a

shorter time-span after fear inducement than control mice), and they have a diminished ability to acquire fear memories (it takes a longer period of shock-training to start behaving fearfully whenever they hear tones).

What epitomizes this experiment is Bickle's (2006, p. 425) maxim: "intervene molecularly and track behaviorally". This is typical of other experiments as well (cf. Parsons and Ressler, 2013 for a review). Bickle (2006, p. 427) goes on to claim that, because such molecular-behavioral correlations obtain, they make the appeal to higher levels (neural networks, cognitive boxology) completely unnecessary, thus supporting his own position of "ruthless reductionism" (p. 428). The reduction is supposed to be that of cognitive function to molecular-behavioral correlations.

While Bickle's attitude is philosophically rather extreme, the pattern he highlights is pervasive in animal models. One could, in light of §2 above, ask where the psychological construct in such animal models for anxiety or depression is. Choi et al. (2012, p. 3) speak of emotional memory formation and consolidation, especially fear memory. But they do not even hint at the possibility of mouse psychology. Fear, the psychological construct, is tacitly equated with fear, the *operational* concept (cf. Sullivan, 2010, p. 878) defined by the experimental protocol of seeing which mouse behaviors are elicited by audible tones and electric shocks. The reasoning perhaps runs as follows: shocks are painful, so the mouse *should* be afraid (in some sense of fear). So, whatever is actually happening with the mouse when it hears a tone after training, experimenters should consider that to be an adequate substitute for fear as characterized in humans.

In terms of psychological constructs, that reasoning spells disaster. The basic emotion of fear differs from both stress (physiological) and avoidance (behavioral). The behavioral reaction of freezing can have a variety of causes: an attention deficit, a stressful environment for a normal perceptual organism, a general state of excitation, etc. Not only are these different psychological constructs, but some of them are even operationally quite discernible from one another. But all count as "fear" for the purposes of Choi et al.'s experiment, instead of just markers of fear. Analogous remarks could be made about Choi et al.'s characterization of mouse addiction to cocaine (a reward-seeking behavior) as repeated place-preference for the chamber where cocaine is distributed – the behavior is clear, the molecular manipulation (same as above, with BDNF) is clear, yet how the psychological construct relates to them is not.

Such *lacunae* in appropriately identifying a psychological construct are in no way specific to Choi et al. (2012). In fact, their study is an excellent one and opens new perspectives from the standpoint of general research in the area. The problem runs deeper. As Choi et al. (2012, p. 6) recognize, their study is intended as a general study of emotional memory that may impact studies of anxiety and depression, but may also be useful for studies of normal fear and reward/satisfaction. What I take the problem to be is that such a difference between (a) individual stimulus-dependent normal psychological constructs, and (b) psychiatric disorders or syndromes such as varieties of anxiety or depression, while it is amply documented in clinical human cases, can hardly be made sense of in animal experimental protocols.

It is not by accident or by lack of experimental precision that animal models like that of Choi et al. cannot capture the difference between (a) and (b). What such experiments achieve is exactly what Bickle praises them for: correlations between molecular interventions and behavioral effects. But if animal models are supposed to be analogical for what happens with humans if molecular intervention is mimicked by pharmacological intervention, then one should expect that something in mice should be analogous to the psychiatric disorders of humans. Current experiments look for that something at a molecular level (neurons and synapses, specific neuromodulators like BDNF), and in specific structures (e.g., the prelimbic cortex). Is this double assumption of localization of target psychological constructs correct?

In human cognitive neuroscience, a popular theoretical framework (the "global network hypothesis", Dehaene et al. 2006, p. 205) runs against the idea that cognitive functions can be localized in a way that is as finely grained as to make contact with molecular interventions. It would follow that one mechanism evidenced in mice (BDNF-TrkB signaling in the prelimbic cortex), while it may be part of the fear-reward pathway that somehow underlies both anxiety and depression, can in *no* way be identified with one of the main determinants of such behaviors.

Non-localization of psychological constructs is also supported by two episodes in the history of neuroscience. The first is that early researchers took the amygdala to be the *locus* of emotional memory processing, whereas now surrounding cortical layers are assigned that function. Progress in research may reveal, in rodents just as in humans, that Dehaene et al. (2006, p. 205) are correct in implying that emotional processing is more widely distributed in the relevant pathways. If so, then the theoretical value of currently established individual molecular-behavioral correlation is questionable, even when those correlations hold and the experiments establishing them have sound controls. Secondly, consider a more recent result, namely, that the reward pathway is regulated the infralimbic rather than the prelimbic cortex (Choi et al., 2012), so it would be improper to assign even the same neural structures to mood disorders like anxiety and depression.

Sullivan (2010, p. 878) makes a similar point against Bickle's reductionism, and similarly remarks on how widely spread implicit operationalist attitudes are among researchers in the field of spatial memory (e.g., Morris et al.'s (1982, p. 683) famous water maze model). Her conclusion is that if progress is to be made in finding molecular correspondents for cognitive functions, there is "a role for representation in cognitive neurobiology" (p. 886). By parity of reasoning, animal model experimentalists for psychiatric disorders should, *contra* Bickle's reductionist recommendation, consider the differences between their intervention protocols and the psychological constructs they wish to track.

4. The analogical argument involved

The suggestions above allow for a more detailed consideration of what animal models are supposed to achieve in relation to psychiatry. Here is an argument pattern that could be instantiated by a successful animal model of a mood disorder:

(1) Rodent behavior X is analogous to human behavior X'.

(2) Rodent behavior X is caused by activation of the neural mechanism Y.

(3) In rodents, the neural mechanism Y realizes/implements the psychological constructZ.

(4) The psychological construct Z in rodents is analogous to the target psychological construct Z' in humans.

Therefore,

(5) In humans, the neural mechanism Y realizes/implements the psychological construct Z'.

A few words about this argument pattern. Notice first that it is an analogical argument: it uses several features that are analogous in rodents and humans, and ascribes a target property to humans on the basis of its obtaining in rodents. The analogical argument can then be challenged on the basis that rodents and humans are actually disanalogous in ways that are relevant to carrying over the target property from rodents to humans. As Bolker (1995, pp. 451–453) remarks, here are two salient differences:

(i) Animal experiments are carried out in an extremely controlled laboratory environment, and this undermines any kind of validity for animal models if conclusions are to be ecologically relevant. If animal models are not even analogous to animals in the wild, how will they be analogous to humans? This is critically important for psychological constructs. For instance, could real-world environmental stress in a war theater and during which soldiers may end up with PTSD be simulated by carefully timed artificial shocks to rodents?

(ii) Animal experiments are carried out with (and animal models rely on) selected animal lineages; such lineages will have been grown in laboratories, and selected for some properties that make them more accessible to research than other lineages. If Bolker (1995, p. 453) is correct, the very reason why it is so practical to use some animal population is because the animals are not representative of their species at large, and so (i) has extra force because of epigenetic changes in the population caused by a non-natural environment.

Both differences should give us pause. But if they were so detrimental, then no animal model would generally be valid (in any sense). However, many animal models in other areas of research (visual system, circulatory system, etc., cf. Chader, 2002, p. 393) are quite successful even though such differences obtain. While there are general concerns about animal models, it may pay off to focus on what is specific to animal models of psychiatric disorders and see if the argument pattern fits them.

The argument pattern above is much weaker than the pattern presented in La Follette and Shanks (1995, p. 147). The extra requirement that they impose is strong: that the neural and behavioral rodent components of the analogy be causally connected

in a known way.[1] It is one thing to say: intervene molecularly and then see what behavioral effects ensue; it is much more to demand that one knows *which* causal chains lead from neural causes to behavioral effects.

A similarly strong line is present in Steel (2010, p. 1060 ff.); the notion of comparative process tracing (in mice and humans) implies that one can trace, in the animal model, the entire process leading up from neural structures to behavioral output. In Steel's terms, what the argument above proposes is a very partial comparative process tracing. In Guala's (2010, pp. 1074-1076) analogical rendering of Steel's work, this amounts to recognizing that the essentially partial character of animal (analogical) models is not in itself an obstacle to their success.

The stronger requirement makes intuitive sense, and fits ambitious research programs; why prefer my weaker version? The weaker version is more in line with current models reviewed in Parsons and Ressler (2013, p. 150). And what we are after is an explanation of why researchers demur from being realists about current animal models, so we need something that addresses the specifics of such models. Consider Choi et al. (2012). They intervene in the prelimbic cortex of mice (p. 4) and observe behavioral effects (p. 5). What the prelimbic cortex-motor behavior causal chains are in the mouse's nervous systems is not a question they address. In trying to understand how their animal model works, a naturalistically prudent course of action is not to overburden models with exacting philosophical standards when models might fail even by practitioners' more permissive standards. Of course, this does not eliminate all causal elements in the account. Premise (2) asserts the existence of a causal connection between neural intervention and behavioral output, and premise (3) asserts that a realization relation obtains between neural activity and psychological constructs that can be understood either causally or, more strongly, constitutively.

A feature of the argument above stands out. Psychological constructs and behavioral effects are supposed to be analogous between rodents and humans. Bracketing the worry that the vast majority of neuroscientists experimenting with animals simply avoid the issue of psychological constructs, what would it mean for rodent behaviors to be analogous to behaviors in humans? Consider another model (Jiang et al., 2010). The authors were looking for evidence about the development of rodent auditory cortex, but one of their key assumptions was that mice would exhibit place-preference for the corner of a square open-field arena that had no music, because music, as opposed to silence, was a stressful environment for rodents (i.e., they could expect danger). How unstressed the mice were was measured by how often they would cross the center square of the arena in exploring the arena and all its corners (music or no music).

[1] La Follette and Shanks do not formulate their condition explicitly. What they say (1995, p. 147) is contraposed: an animal model can be captured by an analogical argument, and any such analogical argument needs to specify how the properties of what does the represented interact causally, if the animal model described is to count as a causal analogical model. Even if my interpretation of their constraint might perhaps be stronger than what they intended, it is worth considering how stronger demands may be placed on animal models than those of (1)-(5) above.

In other words, in Jiang et al. (2010), stress (and anxiety) were operationalized by decreased locomotion in rodents. But anxiety in humans is not confined to decreased locomotion but has important emotional markers. Moreover, it is intuitively very unclear why decreased locomotion, as opposed to locomotor hyperactivity, is supposed to model anxiety better. In claiming that an animal model of anxiety is analogous to its human target, one should clearly distinguish between analogy (they can be put in a 1-1 correspondence that has *some* structural/functional meaning) and similarity (they behave alike). Jiang et al. make the analogy claim, not the similarity one. While perhaps important in capturing relations between other models, theorizing the similarity (Giere, 2004, p. 747) between animal models and their human targets seems not to be the best route. Failure of similarity also questions the usefulness of assimilating the animal model's representation of target systems to a simulation (Piotrowska, 2013, p. 440).

A question is then quite pressing: what exactly is the structural/functional meaning of an analogy between human and rodent behaviors? Given that in rodents the fine-grained distinctions between normal reward and addiction, and normal fear and anxiety, can hardly be made, the analogy in behavior might depend on the existence of analogous psychological constructs (premise 4). That would question the logical independence of premises (1) and (4). While such interrelations are likely to obtain, the argument pattern above seems more appropriate for current research because it points to the need for discussing psychological constructs (premise (4)), while not being oblivious to the experimental emphasis on tracking behavior (premise (1)).

5. Mechanisms and localization

Premise (4) mentions psychological constructs, and §2 above made a case for animal models of psychiatric disorders to include them in any workable model of clinical psychological human conditions. But premises (2) and (3) make reference to a neural mechanism that is supposed to be both causally responsible for behavioral effects and realize the target psychological construct. How should such mechanisms best be understood?

In the model of emotional memory of Choi et al. (2012, pp. 5-6), the synaptic mechanism evinced is that upstream (presynaptic) BDNF is transported to postsynaptic tyrosine kinase (TrkB) receptors. Endocannabinoid retroactive signaling (including TrkB mediated signaling) modulates presynaptic signals to fit demands of the postsynaptic neuron (Castillo et al., 2012, pp. 70-72). This normal mechanism at a synaptic level (for a large set of types of synapses) is disrupted if the BDNF-gene upstream is knocked-down, or if a virus is introduced which depletes prelimbic synapses of BDNF. Notice how, in Choi et al.'s case, the mechanism is specified well beyond the general characterization of entities and activities which produce regular changes from set-up to termination conditions (Machamer et al., 2000, p. 3). The mechanism is clearly localized synaptically at a molecular level, and in the prelimbic cortex at an anatomical-structures level.

This double localization is again typical of advanced current animal models. It also encourages further research, e.g., looking only at glutamatergic synapses instead of

all kinds of synapses (Parsons and Ressler, 2013, p. 150). The mechanistic constraint is in this sense a template of current (good) models. As Zarate et al. (2006, p. 857) suggest, this is in contrast to what happens in the clinical case. Zarate et al. (2006, p. 861) present tests of the effects of ketamine (an NMDA antagonist) on depressed patients. The point relevant here is that Zarate et al. indicate how ketamine fares much better when compared to other antidepressants, and yet the administration of the drug (one injection per week) was site-unspecific. Even more advanced (pharmaceutical) antidepressants, then, lack the kind of localization that animal models of the same disorders do include. In this sense, while animal models may end up playing little justificatory role when all the clinical testing on humans is said and done, they are invaluable hypothetical analogue models (La Follette and Shanks, 1995, p. 159) in the sense that they provide hypotheses about how to chemically intervene at a synaptic level that are more finely grained than what current pharmacological technology can implement.

However, the demand for a psychological construct and the demand for localized molecular mechanisms are somewhat in tension. The issue concerns what Mundale (2002, §3) describes as the physical scope of localization. If cognitive function is widely distributed, then the corresponding mechanisms might have to be enlarged to whole neural pathways. For example, if emotional memory processing cannot be pinned down to any unique structure, it might best be seen as emerging from a fear-reward pathway including the amygdala, the nucleus accumbens, the VTA, and the prelimbic and infralimbic cortices. At that point, if ever reached, it would make little sense to characterize so wide a *congeries* of neural structures performing very different activities as being mechanistic in any discernible sense. Rather than follow Machamer et al. (2000, p. 18) in speaking of hierarchies of mechanistic *schemata*, it might seem more epistemically prudent to confine talk of mechanisms to what happens at individual synapses or neurons, and to classify network-level processes as non-mechanistic in the narrower sense.

Until a deeper understanding of emotional memory as distributed or localized ensues, current animal models function under strong localization assumptions, at both the molecular and anatomical level. For such models like Choi et al., Jiang et al., and those surveyed in Parsons and Ressler, it is not the case that mechanisms in the rodent model are analogous to mechanisms in the human target systems. Rather, in agreement to the argument pattern above (compare (3) and (5)), they wish to claim it is the *same* mechanism at work in the two species. The claim seems plausible at a molecular level, given evolutionary constraints, and a species-invariant characterization of synaptic transmission. Yet the claim depends on the narrow construal of mechanisms suggested in the previous paragraph. If activities in larger-scale structures counted as mechanisms, then failures of homology and developmental differences could undermine the usefulness of animal models to human research.

Given the two requirements of psychological construct and mechanistic neural realization are in relative tension, each may perhaps be relaxed to the benefit of the other. Yet both are important requirements. The operationalist attitude (against the first requirement) implicit in much current work, and the desire to assimilate synaptic-level

processes to network-level processes (against the second requirement) are two temptations that can be resisted. Unfortunately, unless such requirements are considered carefully, Bickle's sharp diagnostic of the situation remains accurate, and all that animal models of psychiatric disorders will manage to produce will be well-documented correlations between some behavioral effects and the molecular interventions that will have triggered them. But an interpretation (a) relating molecular intervention to (lack or rescue of) cognitive function, and (b) relating cognitive function to associated behaviors, would not be forthcoming just from molecular-behavioral correlations.

6. Realism, representation, and intervention

In §§2–5 I have argued that if one is to have a realist attitude towards animal models of psychiatric disorders, these would have to satisfy two requirements: the requirement of psychological constructs and the mechanistic requirement. While there exist partly successful animal models of anxiety and depression that fail to satisfy these requirements (in spite of being predictively valid to some degree), the requirements offer a justification for why researchers typically demur from adopting a realist attitude concerning them.

Suppose that further research did deliver models that satisfied both of the requirements in §§2–5. How should a realist attitude towards such models be construed? Traditionally, one can distinguish between a representational realism (realism towards values of variables in the model), and a realism vis-à-vis entities causally operative in experiments or laboratory interventions (Hacking 1982, p. 72). I would like to suggest that these two aspects (representational and causal-realist) are inextricably interwoven in the case of animal models for mood disorders.

Animal experimentation is one area where the chances of understanding causation in interventionist terms (Menzies and Price, 1993, pp. 187-192) are among the highest. To say, with Choi et al. (2012, pp. 5-6) that BDNF-TrkB signaling in the prelimbic cortex causes formation and consolidation of fear memory, is documented by *intervening* and comparing to controls: what happens when the BDNF-gene is knocked-down, when TrkB-signaling is artificially induced in spite of BDNF absence, etc. Researchers are typically cautious so as not to make any claims of causation *tout court*, but only of necessary or sufficient causal *conditions* (Mackie, 1965, for an example cf. Jiang et al., 2010).

When causal claims about BDNF are made, the existence of BDNF is assumed (both as a gene, and as a protein). But not only is it assumed, but the whole body of knowledge about BDNF in particular, and about neurotrophins and neural modulators in general, is brought to bear on the conclusion from the experiment (when describing the animal model). That body of knowledge is brought to bear in spite of the scarcity of general laws in the life sciences (Machamer et al., 2000, p. 7) because one can still speak of a wider or narrower class of interventions under which a hypothesis is empirically supported (Woodward, 1997, p. S31). So, even though a set of fundamental laws is not characteristic of life sciences, a theory in the sense of an explanation of how pervasive mechanisms work (e.g. synaptic transmission) does exist (Machamer et al., 2000, p. 16). Relative to that body of knowledge, individual animal models are

not autonomous (Morrison 1999, pp. 42–43), but depend on it for their overall validity (to whatever degree it might obtain).

The novelties in a successful animal model of a psychiatric disorder are brought to the fore in the background of accepted theoretical claims about what kinds of interventions may be performed at the molecular level and what kind of thing (either in terms of behavior, or of cognitive function) the experiment tracks. Unlike some cases in fundamental science (Hacking, 1982, p. 74, gives the example of different analogical models of the electron), belief in the existence of entities causally operative in experiments is not divorced from how those entities are theoretically represented.

On the other hand, representation depends on experimentation too. It would be very hard to see how hypotheses about human target systems could be advanced on the basis of animal models *unless* carefully controlled experimentation were performed and replicated with those animal models. The very point of using animal models is that it would *a priori* be very difficult to see what hypotheses about animal models would carry over to their human targets in the absence of the lead of experimentation. This echoes a point made by La Follette and Shanks (1995, p. 159), since future animal models in psychiatry would allow a realist attitude *only if* the hypotheses about humans were derived in ways that evinced an analogy between the causal networks involved in the animal models and their human target systems (albeit I would add that such networks would of practical necessity be partial and focused on how individual mechanisms or pathways might realize target cognitive functions).

7. Conclusion

In this text, I tackled the question of why experimental neuroscientists and psychiatrists alike are not realists about animal models of mood disorders. I explained their reservations by the failure of current models of anxiety and depression to satisfy two requirements that need to be balanced: that of psychological constructs and that of mechanism. This leaves the door open for the possibility of improvement in animal models, which could warrant a less antirealist attitude. I sketched an argument pattern that is operative in characterizing animal models, and I pointed to some limitations in the current methodology for relating molecular manipulations, behavioral tests, and the cognitive functions posited. Finally, I suggested that a distinctive trait of animal experimentation concerns the inseparability of representational and causal-interventionist aspects in the resulting animal models.

Acknowledgments

This work was supported by a grant from the Ministry of Research, Innovation and Digitalization, CNCS-UEFISCDI, project number PN-IV-P2,1-TE-2023-1806, within PNCDI IV. The project supported, titled "An approach to modal epistemology from the standpoint of virtue epistemology and its applications" ("O abordare a epistemologiei modale din perspectiva epistemologiei virtuţilor, și aplicațiile sale"), is hosted by the Research Institute at the University of Bucharest (ICUB, Humanities Division). The initial ideas for this text emerged when I was a doctoral student supported by the Jefferson Scholars Foundation *via* a John S. Lillard fellowship. Subsequent reworking

was based on talks given at the 50^{th} annual meeting of the Long Island Philosophical Society and the 3^{rd} congress of the Société de la Philosophie des Sciences, both in 2014; I'm grateful to the audience and organizers for suggestions, as well as to the late Paul Humphreys (University of Virginia) for immensely helpful comments, as well as a guarded skepticism about my conclusions.

References

Bickle, J. (2006). Reducing mind to molecular pathways: Explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese, 151*, 411–434.

Bolker, J. A. (1995). Model systems in developmental biology. *BioEssays, 17*, 451–455.

Cartwright, N. (1991). Fables and models – I. *Proceedings of the Aristotelian Society, Supplementary Volume, 65*, 55–82.

Castillo, P. E., Younts, T. J., Chávez, A. E., & Hashimotodani, Y. (2012). Endocannabinoid signaling and synaptic function. *Neuron, 76*, 70–81.

Chader, G. J. (2002). Animal models in research on retinal degenerations: Past progress and future hope. *Vision Research, 42*, 393–399.

Choi, D. C., Gourley, S. L., & Ressler, K. J. (2012). Prelimbic BDNF and TrkB signaling regulates consolidation of both appetitive and aversive emotional learning. *Translational Psychiatry, 2*, e205. http://www.ncbi.nlm.nih.gov/pubmed/23250006

Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences, 10*, 204–211.

Giere, R. (2004). How models are used to represent reality. *Philosophy of Science, 71*, 742–752.

Guala, F. (2010). Extrapolation, analogy, and comparative process tracing. *Philosophy of Science, 77*, 1070–1082.

Hacking, I. (1982). Experimentation and scientific realism. *Philosophical Topics, 13*, 71–87.

Jiang, B., Huang, S., de Pasquale, R., Millman, D., Song, L., Lee, H. K., Tsumoto, T., & Kirkwood, A. (2010). The maturation of GABAergic transmission in visual cortex requires endocannabinoid-mediated LTD of inhibitory inputs during a critical period. *Neuron, 66*, 248–259.

La Follette, H., & Shanks, N. (1995). Two models of models in biomedical research. *The Philosophical Quarterly, 45*, 141–160.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1–25.

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly, 2*, 245–264.

Menzies, P., & Price, H. (1993). Causation as a secondary quality. *British Journal for the Philosophy of Science, 44*, 187–203.

Morris, R. G., Garrud, P., Rawlins, J. N., & O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature, 297*, 681–683.

Morrison, M. (1999). Models as autonomous agents. In M. S. Morgan & M. Morrison (Eds.), *Models as mediators* (pp. 38–65). Cambridge University Press.

Mundale, J. (2002). Concepts of localization: Balkanization in the brain. *Brain and Mind, 3*, 313–330.

Parsons, R. G., & Ressler, K. J. (2013). Implications of memory modulation for post-traumatic stress and fear disorders. *Nature Neuroscience, 16*, 146–153.

Piotrowska, M. (2013). From humanized mice to human disease: Guiding extrapolation from model to target. *Biology and Philosophy, 28*, 439–455.

Steel, D., & Kedzie Hall, S. (2010). A new approach to argument by analogy: Extrapolation and chain graphs. *Philosophy of Science, 77*, 1058–1069. https://www.msu.edu/~steel/Analogy_Extrapolation.pdf

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1–25.

Sullivan, J. (2010). A role for representation in cognitive neurobiology. *Philosophy of Science, 77*, 875–887.

Willner, P. (1984). The validity of animal models of depression. *Psychopharmacology, 83*, 1–16.

Woodward, J. (1997). Explanation, invariance, and intervention. Philosophy of Science, 64, S26–S41.

Zarate, C. A., Jr., Singh, J. B., Carlson, P. J., Brutsche, N. E., Ameli, R., Luckenbaugh, D. A., Charney, D. S., & Manji, H. K. (2006). A randomized trial of an N-methyl-D-aspartate antagonist in treatment-resistant major depression. *Archives of General Psychiatry, 63*, 856–864.

Balkan Journal of Philosophy (2026), vol. 18, issue 1, 99 – 115

A MARBLED EPISTEMOLOGY: EMBODIED MODELING AT THE INTERSECTION OF ART AND SCIENCE

Daniela Godoy*

Abstract: This article argues that water marbling affords genuine, non-propositional understanding – tacit, affective, and embodied – thereby strengthening aesthetic cognitivism. Marbling – floating pigments on water, manipulating them into patterns, and transferring them onto paper – renders visible the dynamics of fluid motion, resistance, and emergence. Drawing on philosophical accounts of fictional modeling and aesthetic cognitivism, the article uses select ideas from model-based science (idealization, parameter variation, counterfactual exploration) heuristically to illuminate how marbling fosters understanding through embodied simulation, imagination, and material engagement – without claiming that marbling is scientific modeling or serves scientific aims. Techniques such as ebru and suminagashi are not merely decorative; they foreground analogical, dynamic exploration of complex material behavior. By enabling counterfactual reasoning, perceptual insight, and tactile grasp of uncertainty, marbling exemplifies a "marbled epistemology" – a fluid mode of knowing that softens the perceived art-science divide by clarifying how artistic practice yields understanding. By "divide" here I mean a difference in aims and norms (truth-tracking explanation versus aesthetic, experiential insight), not an ontological wall between domains.

Keywords: epistemic imagination, embodied cognition, fictional modeling, aesthetic cognitivism, hybrid epistemology.

1. Introduction: Unmixing the Great Divide

The long-standing "great divide" between science and art assumes they are governed by fundamentally distinct aims and values: science seeks objective knowledge and truth, while art is primarily associated with aesthetic pleasure and subjective interpretation. This dichotomy, inherited from Enlightenment-era epistemology and Romantic aesthetics, continues to inform both public discourse and institutional structures (Frigg & Nguyen, 2020). Water marbling refers to a family of artistic techniques in which ink or paint is floated on the surface of water or viscous liquid, manipulated into intricate patterns using breath, styluses, or combs, and then transferred onto paper, fabric, or other substrates. The result is a one-of-a-kind image shaped by the interplay of surface tension, pigment behavior, and gestural movement. Unlike conventional painting, where marks are made directly on a surface, marbling composes its forms within a fluid medium – rendering visible the dynamics of flow, resistance, and emergence in real time. This aqueous process is not only visually mesmerizing, but conceptually rich, offering heuristic parallels for understanding fluid systems, embodied cognition, and aesthetic form, while remaining an artistic practice. However,

*Department of Arts and Humanities, American University of Ras Al Khaimah;
t.normagodoy@aurak.ac.ae

recent developments in both the philosophy of science and cognitive aesthetics suggest this bifurcation is not only simplistic, but intellectually untenable.

In the philosophy of science, it is now widely accepted that many scientific models are not literal representations of reality but intentional distortions – idealizations, abstractions, and fictional constructs designed to explore, simplify, or make sense of complex phenomena. As Godfrey-Smith (2006) famously argued, scientific models often behave more like fictions than mirrors; they do not represent the world as it is, but as it might be under certain imagined conditions. More recent accounts by Frigg and Nguyen (2020) elaborate this view by proposing a "fiction view of models," wherein models are treated analogously to literary fictions: as structured imaginings that afford cognitive access to possible states of affairs. In what follows, I borrow these modeling notions analogically to clarify aspects of marbling's practice; they are not used to collapse artistic aims into scientific ones.

On the other side of the traditional divide, a growing body of research in aesthetic cognitivism has shown that engagement with art and fiction often results in forms of understanding and insight, not merely entertainment or aesthetic pleasure. Gaut (2003) and Currie (2020), among others, argue that narrative fiction, visual art, and even poetry can function as epistemic tools: they foster analogical thinking, simulate experiences, and deepen our grasp of emotional, psychological, or social truths. Marmodoro (2020) further suggests that art enables what she calls "non-propositional cognition" – ways of knowing grounded in perception, feeling, and embodied interaction. Freedberg and Gallese (2007) build on this by showing that artworks can elicit embodied and empathic responses, while Leder et al. (2004) demonstrate how even abstract, non-narrative art prompts imaginative and affective engagement. Keen (2007) echoes this view, emphasizing that fiction and art do not merely convey facts but cultivate forms of psychological insight, empathy, and self-understanding. At the same time, as Currie notes, such imaginative engagement is epistemically powerful yet fallible; my claim in this paper is therefore domain-bound to perceptual grasp, material affordances, and self-regulation within practice.

Despite these developments, discussions around scientific and artistic understanding have largely remained disciplinarily siloed. Only a few scholars (e.g., Bueno et al., 2017) have attempted a serious comparison between the epistemic mechanisms at work in the two domains. The result is a persistent epistemological blind spot: while science and art may share tools, techniques, and aims, our intellectual frameworks still treat them as fundamentally distinct.

This paper proposes that marbling – the embodied artistic practice of floating pigments on water and transferring their patterns to paper–offers a powerful case study to rethink this binary. Historically practiced in Japan (suminagashi) and Turkey (ebru), marbling is both a meditative ritual and a material exploration of dynamic systems. It engages physical properties such as viscosity, surface tension, and fluid motion in a controlled but unpredictable way. Rather than equating marbling with scientific modeling, I argue that marbling can be clarified by concepts used in model-based science (idealization, structured variation, counterfactual exploration), because marbling likewise works through intentionally unstable, abstract, and analogical scenarios. At the

same time, marbling generates aesthetic meaning and affective response. It is not mere data visualization; it is sensuous thinking through materials.

The central question this paper asks is:

> *Can marbling function as an epistemic artistic practice that yields understanding without propositional claims?*

I argue that it can. Through a close reading of marbling as embodied, affective, and analogical practice – illuminated heuristically by concepts from model-based science and by aesthetic cognitivism – I show how it softens the perceived divide between "truth" and "beauty," "fact" and "imagination." Marbling cultivates insight – into systems, into self, into emergence and impermanence – without adopting scientific aims or inferential norms.

This paper proceeds in six sections. Section 2 explores the epistemic status of scientific models, focusing on their fictional and heuristic nature. Section 2 serves as background only; its modeling notions are used analogically rather than as identity claims. Section 3 provides a phenomenological account of marbling as an artistic and epistemic activity. Section 4 engages with theories of aesthetic cognitivism to frame marbling as a form of experiential understanding. Section 5 discusses the role of ekphrasis – writing about visual experience – as a way of translating marbled practice into conceptual reflection. Section 6 analyses the role of convention in enabling both artistic and scientific insight. I conclude by proposing a "marbled epistemology" that holds promise for future cross-disciplinary conversations about art, science, and the shared pursuit of understanding.

2. Scientific Models as Fictional Tools

Philosophers of science have increasingly recognized that scientific models often rely on deliberate distortion and idealization, functioning less like mirrors of reality and more like idealized constructs designed to facilitate understanding (Godfrey Smith, 2006). Peter Godfrey Smith presents a vivid case for this "model-based" approach: scientists craft simplified, hypothetical systems – rabbit populations without death, frictionless planes, ideal gases – not to assert their literal existence, but to explore patterns, relationships, and dynamics in a controlled conceptual space (Godfrey Smith, 2006). He observes that these models behave similarly to fictional domains in literature: invented, counterfactual worlds with their own internal logic (Godfrey Smith, 2006).

Expanding on this idea, Frigg and Nguyen (2020) argue that scientific models function analogously to literary fictions. In their "fiction view," models are crafted imaginings used to reason about possibilities, enabling scientists to draw inferences about how systems could behave (Frigg & Nguyen, 2020). Their DEKI account of representation further illuminates this dynamic: models denote target systems, exemplify relevant features, are keyed up via translation protocols, and then imputed onto real-world phenomena (Frigg & Nguyen, 2020).

Crucially, model-based reasoning supports non-factive understanding. Authors like de Regt (2017) and Khalifa (2017) have argued that understanding often does not require strictly true premises. Instead, familiarity with how variables interact within a structured framework can yield insight – even if the model itself is idealized or simplified. For example, de Regt (2017) introduces the concept of "grasping" explanatory patterns, showing that understanding isn't reducible to truth-tracking alone. Khalifa (2017), likewise, emphasizes that "counterfactual understanding" – knowing how variations would play out – carries significant epistemic weight.

This philosophical stance aligns with Godfrey-Smith's example of the frictionless plane: although it misrepresents real-world physics, it still contributes to knowledge about motion dynamics. It allows one to understand the structure and consequences of Newton's laws within a hypothetical yet illuminating ideal case (Godfrey Smith, 2006). Thus, modeling scientific understanding as story-like reasoning, rather than propositional mapping, serves cognitive advancement – even amid fictionalization.

Why describe scientific models as fictional? The key philosophical insight is that fictional models are objects of imagination. They enable scientists to simulate, experiment, and reason within possible worlds – contexts shaped by abstraction, simplification, or idealization. Like a novel or thought experiment, a model allows creators and readers to explore what-if scenarios, test hypotheses, understand dependencies, and cultivate counterfactual reasoning (Frigg & Nguyen, 2020).

This perspective has crucial implications for drawing parallels between art and science. If scientific models are structured fictions whose primary function is to cultivate understanding – analogous to the imaginative work of literature – then the supposed gulf between objectivity and subjectivity, fact and fiction, begins to break down. Science, in this view, is deeply imaginative, while art, too, constructs scenarios that foster insight.

These insights ground the comparison to marbling: the practice creates controlled yet unstable material fictions – blended pigments that simulate fluid dynamics, emergent behaviors, or stochastic movement. Much like scientific models, marbling requires manipulating variables and observing outcomes, forming embodied analogies and patterns that resonate with viewers' intuitions and percepts. It invites counterfactual exploration ("what if I drop more ink?" "What if I stir faster?") and pattern recognition – all without propositional declarations. Accordingly, when I draw on modeling vocabulary, I use it heuristically: marbling exhibits modeling-like operations (idealization, parameter variation, counterfactual play) in a practice-bound, artistic context. It cultivates non-propositional understanding in a neighboring region of exploratory, counterfactual, and pattern-grasping practice to that inhabited by scientific modelers, while differing in aims, norms, and outputs.

3. Marbling as embodied epistemic practice

The traditional narrative of water marbling often emphasizes its visual elegance, decorative role, or cultural symbolism. However, when viewed through the lens of embodied cognition and material engagement, marbling emerges not merely as an

aesthetic activity, but as an epistemic practice – one that enacts knowledge through dynamic interaction with matter. This section traces how the embodied process of marbling models the logic of scientific experimentation, while simultaneously producing perceptual and analogical understanding.

3.1. Historical and material foundations

Water marbling traditions such as suminagashi (Japan) and ebru (Ottoman Empire) have been carefully preserved and passed down through generations, primarily as artisanal techniques used for book decoration and manuscript margins (Loring, 1942). While Japanese suminagashi traditions emphasized meditative and minimalistic approaches using ink on plain water, Turkish ebru developed more systematic approaches involving thickened water mediums and specialized tools such as combs and styluses for pattern manipulation. This evolution from contemplative practice to systematic manipulation represents a significant development in decorative arts: marbling transformed from passive observation to active procedural control. The marbler's practice evolved to involve real-time manipulation of fluid dynamics, creating what can be understood as an analogical system where specific variables – including medium viscosity, pigment properties, and tool pressure – are systematically altered to generate controlled structural transformations (cf. Weisberg, 2013). This helps explain why marbling specifically matters epistemically: on-water formation with high sensitivity to initial conditions, tight parameterization (viscosity, surfactant, drop size, tool spacing), immediate perceptual feedback, and codified pattern grammars (battal, gelgit, taraklı) together foreground reliable counterfactual skill and fine-grained tacit calibration in ways many other art forms do not. Jaffer's (2018) mathematical analysis of marbling confirms this. Using differential equations, Jaffer demonstrates that many marbling techniques – including the creation of gelgit (back-and-forth) or şal (shawl) patterns – can be precisely described as homeomorphisms of incompressible Newtonian fluids. In short, marbling is not simply expressive; it models fluid behavior through material abstraction and stylized distortion, much like scientific simulations of turbulence or wave propagation (Jaffer, 2018).

3.2. Phenomenology and embodied knowledge

While mathematical modeling provides an external account of marbling, phenomenology offers an internal one. The marbler does not approach the tray as a disembodied technician, but as an affective, attuned body. The ink does not simply obey mechanical laws; it responds to breath, rhythm, temperature, intuition. Maurice Merleau-Ponty (1945/2010) emphasizes that perception is never purely intellectual; it is "motor intentionality," grounded in the pre-reflective body's orientation toward the world. The hand that releases the ink, the eye that tracks its spread, and the breath that stirs the surface form a living system of perception.

Artistic research has echoed this insight. Banfield and Burgess (2013), in their phenomenological study of professional artists, argue that artistic flow is not merely an affective state, but a form of embodied knowing. Artists report "thinking through the

material," where meaning arises not from planning, but from responsive engagement with surface, resistance, and movement (Banfield & Burgess, 2013, p. 3). Similarly, in marbling, understanding emerges not from preconceived intention, but from adjusting to the ink's shifting behavior: if the pigment disperses too quickly, a marbler alters the ratio; if it resists a comb, she modifies pressure.

This aligns with what Polanyi (1966) calls "tacit knowledge" – knowledge that we know how to use without being able to articulate. The marbler's mastery is not reducible to a formula. It is enacted in time, through sensation, repetition, and material memory. In this sense, the marbler's expertise exemplifies Polanyi's tacit dimension: situational grasp that is actionable, teachable by demonstration, and reliably sensitive to parameter change.

3.3. The tray as laboratory

If we understand scientific models as fictional, analogical systems for exploring abstract patterns (Godfrey-Smith, 2006; Frigg & Nguyen, 2020), then marbling can be clarified as a kind of embodied exploratory system that shares modeling-like operations without scientific aims. The marbling tray is a microcosm – a dynamic field governed by change, constraint, and emergence. The practitioner adjusts one variable (drop size), observes its effect on another (spread diameter), and iterates. The result is a phenomenological experiment: not one that produces universal truths, but one that reveals structures of becoming.

For example, when a pigment is floated on thickened water, its dispersion is shaped by surface tension, pigment density, water temperature, and the marbler's delivery technique. Each action becomes a test. The comb, in this context, operates like a tool of systemic intervention – a gesture that transforms one configuration into another. As such, marbling mirrors the structure of a parametric model, in which initial conditions and transformation rules generate a range of outcomes. But unlike traditional models which isolate variables for clarity, marbling's power lies in its non-linearity and aesthetic opacity. The marbled pattern cannot be fully predicted or stabilized; it resists closure. This indeterminacy is not a flaw. It is a feature that aligns marbling with systems thinking, emergence, and complexity – concepts increasingly important in scientific modeling today (Mitchell, 2009).

3.4. From process to insight

While marbling may not offer propositional knowledge in the traditional sense, it facilitates epistemic insight through material resonance. To marble is to think-with-water, to experiment-with-flow, to feel-with-color. The artist does not merely see results; she intuits the logic behind them. This is why marbling, like scientific modeling, allows for counterfactual reasoning: What happens if I introduce a heavier pigment? What if I tilt the tray? What if I change the comb interval? These are not aesthetic choices alone—they are modeling-like exploratory questions about parameters and outcomes. The marbler, in other words, is not just decorating paper; she engages in structured exploratory variation akin to model-play. Marbling, when viewed through

the lenses of embodied phenomenology and model-based reasoning, emerges as a profound epistemic art form. It is not just beautiful; it is cognitively rich. It does not merely represent the world; it models its fluidity, emergence, and contingency. In the hands of the marbler, the tray functions as a disciplined site of exploratory variation, the ink a medium that invites hypothesis-testing in a loose, practice-bound sense. As such, marbling deserves a central place in philosophical discussions about the epistemic potential of art.

4. Aesthetic cognitivism and the epistemic role of marbling

Recent work in aesthetic cognitivism challenges the notion that art is solely for entertainment or aesthetic pleasure. Instead, art can be epistemically valuable, fostering understanding, insight, and even emotional intelligence (Currie, 2020; Gaut, 2003). In this section, I argue that marbling functions as an epistemic practice not just through material modeling (as shown in Section 3), but by engaging viewers – and practitioners – in cognitive processes historically associated with both art and science.

4.1. Art, fiction, and understanding

Gaut (2003) posits that art can convey knowledge through non-propositional means, such as sensory engagement, allegory, and visual metaphor. Currie (2020) extends this by suggesting that narrative fiction allows readers to simulate social realities, fostering moral and psychological insight. Marmodoro (2020) goes even further, describing art as capable of generating non-conceptual forms of understanding, rooted in perception and imaginative simulation. Keen (2007) specifically emphasizes the role of narrative empathy – the ability to internalize another's perspective – as central to art's epistemic value. By enabling counterfactual and analogical thinking, art resembles the fictional modelling seen in science, yet it achieves insight through different cognitive pathways – particularly affective and embodied routes (Currie, 2020; Marmodoro, 2020).

4.2. Marbling as a cognitive simulator

Marbling is not a traditional narrative form, but it shares essential characteristics with both scientific and literary imagination. The process invites analogical reasoning ("ink behaves like a fluid," "patterns resemble cellular structures," "waves of color evoke emotion"), and supports counterfactual exploration ("What if I change viscosity?", "What if I introduce combing in circles?"). These are precisely the forms of thought touted by aesthetic cognitivists (Currie, 2020; Marmodoro, 2020; Freedberg & Gallese, 2007; Leder et al., 2004).

Moreover, cognitive science shows that viewers can extract psychological insight even from non-narrative artworks. Research in neuroaesthetics supports this: Freedberg and Gallese (2007) argue that abstract forms evoke embodied empathy through motor resonance, while Leder et al. (2004) demonstrate how affective syntax – the interplay of form, color, and movement – generates cognitive and emotional insight.

This aligns with the way marbling patterns can evoke moods, feelings, and embodied responses (e.g., calming, turbulent, meditative), generating understanding without declarative propositions. At the same time, as Currie emphasizes, such imaginative engagement is epistemically potent yet fallible; here I restrict the claim to practice-bound perceptual grasp, affordance sensitivity, and self-regulation rather than generalizable propositions.

4.3. Ekphrasis meets embodied representation

Ekphrasis – descriptive writing inspired by visual art – has traditionally been seen as secondary or decorative. However, in recent scholarship it's gaining recognition as an artistic and philosophical tool in its own right, capable of carrying insight (Heffernan, 1993; Sontag, 1966). By writing ekphrastically about marbling, the practitioner-poet bridges embodied experience and conceptual reflection: how something looks becomes entwined with what it discloses about systems, sensations, and the limits of representation. Ekphrastic reflection on marbling uncovers what it's like to mar – the tactile choices, the mineral resistance, the flow of color – thus transforming embodied, pre-conceptual experience into shared insight. This process shows how marbling, like narrative art, can lead to forms of mutual understanding – both of external phenomena and of the self.

4.4. Insights, empathy, and self-knowledge

Artistic practices such as marbling do more than simulate external systems; they probe the artist's own psychology. The fleeting swirls and balance of chaos lend themselves to reflective metaphors about life's impermanence, emotional complexity, and relational presence – much as narrative fiction prompts self-reflection (Currie, 2020). The marbler, through subtle calibrations, learns about patience, spontaneity, failure, and acceptance – much as narrative fiction prompts self-reflection (Keen, 2007).

In this way, marbling fosters what Schindler (2015) terms embodied tacit knowledge – awareness of one's tendencies, constraints, and habits through hands-on engagement with materials (Schindler, 2015). This form of self-knowledge is epistemically valuable – not merely anecdotal – but foundational to creative and scientific insight.

4.5. Synthesis: marbling as hybrid epistemic practice

By combining material modeling (Section 3) with cognitive and affective engagement (as shown here), marbling emerges as a unique form of hybrid epistemic practice. It shares with science:

- model-based reasoning through controlled variation,
- emergent systems thinking, and
- counterfactual simulation, and with art:
- embodied cognition,

- analogical resonance,
- narrative (or aesthetic) empathy,
- self-reflective insight.

These do not sit on opposite ends of a spectrum but intertwine in practices like marbling. To marble is not "just art" and certainly not contrary to scientific thinking – it is a hands-on convergence of imagination, matter, and meaning.

5. Ekphrasis as witnessing: writing with, not about, marbling

In the classical tradition, ekphrasis referred to vivid verbal description of visual scenes, prized for its rhetorical force and ability to "make present" what was absent (Heffernan, 1993). However, modern and postmodern thinkers have challenged the assumption that language can ever truly "capture" the visual. Instead, ekphrasis is increasingly viewed as a site of tension – between showing and saying, material presence and conceptual grasp (Mitchell, 1994).

In the context of marbling, ekphrasis takes on a new, intensified difficulty. Marbled surfaces resist indexical representation: they are non-linear, multi-centered, and emergent. They do not offer a stable object to describe, but a swirl of process and residue. To write about marbling, then, is not to "represent" it but to enter into relation with it – to write with the image as a co-constitutive event. Ekphrasis thus stabilizes and communicates elements of tacit, pre-conceptual grasp, making them available for inspection, teaching, and critique.

5.1. From description to ontological listening

To describe a marbled surface is already to misrepresent it. The ink's movement, the weight of the brush, the breath that stirs the water – these are untranslatable gestures, ephemeral acts. The final print is not the art itself, but its trace: a frozen echo of an interaction that no longer exists.

In this sense, ekphrasis becomes not description but testimony – a speaking of the experience of being-with the material. This echoes Adriana Cavarero's (2005) argument that voice is irreducible to logos: to speak is not merely to convey information, but to emanate presence. Similarly, to write ekphrastically about marbling is to voice the silent logic of the medium, to acknowledge that the water and the ink "speak" in forms that language cannot replicate.

Cavarero insists on the uniqueness of voice – its timbre, texture, and temporality – as a mode of singular embodiment. Marbling, too, carries this singularity. Each pattern, though built from conventional strokes and tools, contains irreproducible gestures, micro-variations shaped by time, hand, and emotion. Ekphrastic writing, then, becomes a mode of attunement – not capture, but correspondence.

5.2. Irigaray and the ethics of proximity

Luce Irigaray (1993) speaks of an ethics of "being two," where relation is not based on mastery or comprehension, but on respectful proximity. To relate ethically is to resist the temptation to subsume the other into the self, to allow for alterity. Writing about marbling, in this sense, requires a poetics of restraint: a willingness to let the image retain its opacity, to speak obliquely, to whisper rather than define.

Irigaray's philosophy resonates with the marbler's stance: one does not dominate the ink or command the water. One negotiates, responds, listens. The practice itself is dialogical – a back-and-forth, a hovering near. Ekphrastic writing, at its most ethical, must mirror this posture. It must not seek to fix the image in meaning, but to let meaning flicker across the surface, ephemeral as light on a wave.

5.3. Phenomenology of the trace

In Merleau-Ponty's later writings on painting, particularly in Eye and Mind (1964), he emphasizes that vision is not simply the reception of images but a form of tactile being-in-the-world. The painter, like the marbler, does not stand outside the image but is immersed in its formation. The hand does not execute a plan; it follows a resonance.

Marbling is the purest expression of this idea: the pattern is not drawn but discovered. The marbler follows the ink's behavior, responds to its expansions and resistances, adjusts mid-breath. The result is not a planned composition but a trace of perception itself.

To write ekphrastically, then, is not to describe a finished object but to trace a trace – to engage with the echo of the echo. This demands a form of phenomenological writing that mirrors the instability, fluidity, and temporality of the image. It requires that the writer too, like the marbler, surrender to the medium's logic.

5.4. Ekphrasis as co-presence

Instead of treating the visual and verbal as separate modes, we might follow James Elkins (1999), who suggests that writing and image contaminate each other – that each is always already implicated in the other. Ekphrasis, in this view, becomes a hybrid zone, a site where perception and articulation collide. Marbling pushes this idea further: because its images are already non-symbolic and fluid, writing must mimic that movement – must stutter, spiral, dissolve.

Consider the following attempt at ekphrasis: "The blue was not blue until the yellow haloed it. The comb did not draw a line but a question. The paper drank what the water could not hold. I do not describe it. I remember it."

Here, the writing enacts its subject. It hesitates, unfolds, moves like ink. This is not a metaphor for marbling. It is an extension of the marbling process into language.

5.5. Toward a marbled poetics

What would it mean to practice ekphrasis as a marbled poetics – one that embraces fluidity, emergence, and resonance? It would mean letting the artwork shape the rhythm of the writing. It would mean writing not from the outside, but from within the process. It would mean resisting the urge to conclude or define, and instead allowing insight to arise from movement and touch. Such a poetics would be grounded in humility and wonder – hallmarks of both artistic and scientific practice. It would accept that not all understanding can be articulated, and that sometimes the most honest thing we can write is a near-silence.

Ekphrasis, when practiced as philosophical witnessing rather than descriptive translation, becomes a powerful tool for articulating the inarticulable. In marbling, it becomes a means of staying-with the image, of voicing what cannot be captured, of letting the trace remain a trace. Grounded in the philosophies of Cavarero, Irigaray, and Merleau-Ponty, this section has shown that to write ekphrastically about marbling is to engage in an ethical, embodied, and ontological practice – one that bridges not only art and science, but self and other, language and form. In short, ekphrastic practice externalizes embodied insight so that it can circulate as shareable understanding without reducing it to propositions.

6. Conventions, constraints, and fictions: understanding through form

A common assumption in both scientific and aesthetic epistemology is that understanding emerges from freedom: from creativity, innovation, or deviation from the norm. But as scholars in both fields have shown, conventions and constraints are not limits to understanding – they are its scaffolding. Scientific models depend on idealized constraints (e.g., frictionless planes, closed systems); artistic forms rely on traditions, genres, materials, and stylistic grammars.

This section explores how marbling operates through conventions and material constraints, and how those constraints enable epistemic insight – not unlike models in science and narrative structures in fiction. By examining how marbling uses historical forms (gelgit, şal, taraklı, battal) as cognitive tools, we see how constraint becomes a condition for discovery, not an obstacle to it.

6.1. Constraints as cognitive enablers

In philosophy of science, it is well established that idealization is not a flaw but a feature of model construction. Scientific understanding often depends on simplified or distorted versions of reality – what Michael Weisberg (2013) calls "model-worlds" – to isolate variables, generate predictions, or test relationships. Similarly, Frigg and Nguyen (2020) argue that these models are governed by internal conventions that allow scientists to reason "within" the model without assuming it reflects the world in a literal sense.

In fiction, the same is true. Narratives follow genre conventions, symbolic logics, and narrative arcs – not to mirror reality, but to shape meaning within plausible fictions. As Currie (2020) argues, these structures allow readers to engage in simulative reasoning and imaginative projection. Without such form, fiction becomes incoherent.

6.2. Marbling patterns as epistemic conventions

Marbling, too, is steeped in convention. Ottoman ebru artists classified their work into recognizable formal types:

Battal (stone pattern)	floating colors manipulated with minimal intervention, resembling natural randomness.
Gelgit (to-and-fro)	achieved by combing the surface in alternating directions.
Taraklı (combed)	structured waves with equal spacing.
Hafif (light) or şal (shawl) patterns	asymmetrical compositions evocative of textile draping.

These patterns are not merely decorative. They are models—each a system of constraints that reveals something about fluid behavior, resistance, and material response. The repetition of these forms serves as a kind of controlled experimentation: when conditions shift (temperature, viscosity, pigment density), deviations from the conventional form become data that the marbler interprets.

The marbler learns not by reinventing the form each time, but by inhabiting it repeatedly, noticing how the image evolves within constraint. This echoes how scientists use known models to test hypotheses through incremental variation.

6.3. Fictional Frameworks in Scientific and Artistic Understanding

The idea that fictional constructs can yield real insight is central to both model-based science and aesthetic cognitivism. As shown in earlier sections, scientists use intentionally distorted systems to test theories, while artists use formal systems to communicate emotional, social, or epistemic truths.

Marbling aligns with this hybrid tradition: it operates as both a formal structure (a repeatable, generative grammar) and a fictional domain (a world-in-water that simulates motion, rhythm, and emergence). It does not depict real rivers or cosmic flows, but it evokes them. It is not data, but it models dynamic processes in a materially intuitive way. This performative fiction – this aesthetic system governed by formal convention – is what allows marbling to teach, to offer understanding, even in the absence of explicit propositional statements.

6.4. The freedom of convention

Crucially, the existence of constraints does not restrict imagination. It gives it form. As Irigaray (1993) and Cavarero (2005) both note in different contexts, relational presence is not dissolved by limits; it is made possible by them. The space between self and other, image and writer, or ink and water is structured – not erased – by boundaries.

In marbling, the tray imposes a limit. The surface tension sets parameters. The pigments are constrained by chemical properties. But within these conditions, freedom flourishes. The artist explores variation, repetition, mutation. The form becomes a resonant field through which insight emerges. In this sense, marbling teaches an important epistemological lesson: understanding arises not from chaos, but from structured play – from a fluid dance between convention and improvisation.

6.5. Toward a liquid epistemology

If we accept that conventions function as cognitive tools – whether in scientific modeling or artistic form – then marbling stands as a compelling example of liquid epistemology: a way of knowing grounded in flow, constraint, repetition, and touch. It clarifies not only what we come to grasp, but how we come to grasp – through iterative engagement with material systems, aesthetic feedback loops, and embodied intuition shaped by tradition. Unlike rigid systems of logic, marbling's conventions are porous and rhythmic. They do not prescribe outcomes but invite relational emergence. This distinguishes marbling from algorithmic design or propositional models: it is governed, but not deterministic. Structured, but not fixed. Fictional, but not false.

By inhabiting historical forms and practicing within structured constraints, marbling reveals the epistemic power of convention. Like scientific models or narrative structures, marbling patterns serve as fictional frameworks that allow for analogical reasoning, counterfactual thinking, and embodied experimentation. In doing so, marbling resists the myth that understanding arises only from freedom or abstraction. It shows that constraint, repetition, and form are not antithetical to insight – they are its very medium.

7. Conclusion: toward a marbled epistemology

This article began with a provocation: that the entrenched "great divide" between science and art is not only philosophically overstated, but epistemically counterproductive. As disciplinary boundaries ossify around assumptions – science as truth-seeking, art as pleasure-producing – we risk ignoring the nuanced ways in which understanding emerges from practices that are at once material, imaginative, and structured. Through a close examination of water marbling –specifically suminagashi and ebru – this article has proposed a reframing: that marbling is not simply an aesthetic artifact, but an epistemic artistic practice that softens the perceived binary between knowing and sensing, abstraction and intuition – using modeling notions heuristically without collapsing aims or norms.

7.1. Reweaving the argument

In Section 2, we traced how philosophers of science have come to accept that many scientific models are fictional, idealized constructs designed not to mirror the world but to simulate and understand it. Scholars such as Godfrey-Smith (2006), Frigg and Nguyen (2020), and Weisberg (2013) demonstrate that models gain their epistemic value not through truth, but through controlled abstraction and structured imagination. These fictions are functional: they isolate variables, generate hypotheses, and invite counterfactual exploration.

In Section 3, we showed that marbling, too, operates through similar principles. The marbling tray becomes a microcosm of emergent systems, where gesture, viscosity, pigment, and tool interact dynamically. As Jaffer (2018) mathematically demonstrates, marbling patterns can be modeled as homeomorphisms in fluid systems. But more than a physical phenomenon, marbling is also phenomenological: an embodied encounter with process, rhythm, and material resistance. Through Merleau-Ponty's (1945/2010) lens, we see that marbling is a lived practice of attention and intentionality – knowledge enacted through the body.

In Section 4, aesthetic cognitivism offered another bridge. Marbling yields insight not through argument, but through analogical, affective, and perceptual means. Like narrative fiction, marbling invites viewers to simulate possibilities, feel-with the pattern, and experience emergence aesthetically. As Gaut (2003), Currie (2020), and Marmodoro (2020) show, art can generate understanding through non-propositional reasoning – through form, mood, and metaphor. Marbling activates this capacity by offering patterned fictions of motion and transformation.

In Section 5, we turned to ekphrasis as a mode of philosophical witnessing. Drawing on Cavarero (2005), Irigaray (1993), and Merleau-Ponty (1964), we argued that writing about marbling requires not representation, but resonance. Ekphrasis becomes a vocal trace of the material trace. To write with marbling is not to describe it, but to dwell with its ambiguity, to listen to what cannot be said directly. In this way, marbling discloses an ontology of presence, absence, and relation – an ethics of attunement through aesthetic practice.

Finally, in Section 6, we emphasized the role of convention and constraint in epistemic activity. Just as scientific models operate within idealized systems, and literary fiction within genre logics, marbling relies on patterned forms – battal, gelgit, şal – to generate and stabilize variation. These conventions do not restrict imagination; they enable structured play. They become the very medium through which understanding arises.

7.2. What marbling teaches us about understanding

The cumulative insight of this inquiry is that marbling constitutes a form of epistemic practice that is neither exclusively scientific nor artistic, but profoundly both. It teaches us that:

- Understanding does not require propositional knowledge; it can arise from perceptual familiarity, embodied iteration, and analogical resonance.
- Fiction is not the opposite of knowledge, but a generative frame for exploring possibilities.
- Constraints are not epistemic limitations, but the very conditions for discovery.
- Embodied practices matter: the hand, the breath, the weight of pigment – all play a cognitive role.
- Writing about art is itself a form of knowing: ekphrasis is not representational but relational, echoing the logic of the artwork.

Marbling enacts what might be called a liquid epistemology: one that resists stasis, embraces emergence, and moves between categories. It shows that knowing is not always a matter of fixity, but often of attunement to flux – to the shimmer of pattern on water, the flicker of insight in metaphor, the whisper of relation in material contact.

7.3. Implications for the philosophy of art and science

This rethinking of marbling's epistemic status contributes to ongoing philosophical conversations in several ways:

- In aesthetics, it supports the claims of aesthetic cognitivism by offering a concrete, non-narrative example of how art facilitates understanding. Marbling models the logic of emergence, a concept difficult to grasp in theory but intuitively legible in form.
- In philosophy of science, it bolsters the fictionalist view of models by showing that analogical, non-propositional fictions can be materially instantiated and cognitively rich. Marbling shows how idealization, constraint, and variation can function in a non-propositional, aesthetic mode of understanding.
- In epistemology, it challenges the privileging of discursive knowledge over tacit or embodied knowing. Drawing on Polanyi's (1966) tacit dimension and Merleau-Ponty's (1964) notion of the visible-invisible, marbling becomes an exemplar of how we come to know through doing.
- In feminist and post-structuralist thought, it echoes arguments by Irigaray (1993) and Cavarero (2005) about the need to rethink relationality, voice, and non-mastery in knowledge production. Marbling teaches us not to dominate materials but to collaborate with them – to listen, yield, and respond.

7.4. Final reflection: what is a marbled epistemology?

A marbled epistemology is not a metaphor. It is a real, embodied framework for understanding. It resists rigid categories and instead favors layered knowing – knowing that is iterative, patterned, gestural, and fluid. It suggests that the future of philosophical inquiry lies not in defending boundaries between art and science, but in creating practices that traverse them. To marble is to know differently: to enter a

space where precision and intuition coexist, where fiction and materiality are allies, and where language becomes waterborne. In the end, marbling does not merely teach us about art or science – it teaches us about how to understand at all.

References

Banfield, J., & Burgess, M. (2013). A phenomenology of artistic doing: Flow as embodied knowing in 2D and 3D professional artists. *Journal of Phenomenological Psychology, 44*(1), 60–91. https://doi.org/10.1163/15691624-12341245

Bueno, O., Darby, G., French, S., & Rickles, D. (Eds.). (2017). *Thinking about science, reflecting on art: Bringing aesthetics and philosophy of science together*. Routledge. https://doi.org/10.4324/9781315114927

Cavarero, A. (2005). *For more than one voice: Toward a philosophy of vocal expression* (P. A. Kottman, Trans.). Stanford University Press. https://doi.org/10.1515/9780804767309

Currie, G. (2020). *Imagining and knowing: The shape of fiction*. Oxford University Press. https://doi.org/10.1093/oso/9780199656615.001.0001

de Regt, H. W. (2017). *Understanding scientific understanding*. Oxford University Press. https://doi.org/10.1093/oso/9780190652913.001.0001

Elkins, J. (1999). *The domain of images*. Cornell University Press. https://doi.org/10.7591/9781501723902

Freedberg, D.,& Gallese, V. (2007). Motion, emotion and empathy in esthetic experience. *Trends in Cognitive Sciences, 11*(5), 197–203. https://doi.org/10.1016/j.tics.2007.02.003

Frigg, R., & Nguyen, J. (2020). *Modelling nature: An opinionated introduction to scientific representation*. Springer. https://doi.org/10.1007/978-3-030-45153-0

Gaut, B. (2003). Art and knowledge. In J. Levinson (Ed.), *The Oxford handbook of aesthetics* (pp. 436–450). Oxford University Press.

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology & Philosophy, 21*(5), 725–740. https://doi.org/10.1007/s10539-006-9054-6

Heffernan, J. A. W. (1993). *Museum of words: The poetics of ekphrasis from Homer to Ashbery*. University of Chicago Press.

Irigaray, L. (1993). *An ethics of sexual difference* (C. Burke & G. C. Gill, Trans.). Cornell University Press.

Jaffer, A. (2018). The Lamb–Oseen vortex and paint marbling. *arXiv preprint*. https://doi.org/10.48550/arXiv.1810.04646

Keen, S. (2007). *Empathy and the novel*. Oxford University Press.

Khalifa, K. (2017). *Understanding, explanation, and scientific knowledge*. Cambridge University Press. https://doi.org/10.1017/9781108164276

Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology, 95*(4), 489–508. https://doi.org/10.1348/0007126042369811

Loring, R. B. (1942). *Decorated book papers: Being an account of their designs and fashions*. Harvard University Press.

Marmodoro, A. (2020). Aesthetic cognitivism. *JoLMA: The Journal for the Philosophy of Language, Mind and the Arts, 1*(1), 41–52. https://doi.org/10.30687/Jolma//2020/01/003

Merleau-Ponty, M. (1964). Eye and mind. In J. Edie (Ed.), *The primacy of perception* (pp. 159–190). Northwestern University Press.

Merleau-Ponty, M. (2010). *Phenomenology of perception* (D. Landes, Trans.; 1st ed.). Routledge. https://doi.org/10.4324/9780203720714

Mitchell, M. (2009). *Complexity: A guided tour*. Oxford University Press. https://doi.org/10.1093/oso/9780195124415.001.0001

Mitchell, W. J. T. (1994). *Picture theory: Essays on verbal and visual representation*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226245904.001.0001

Polanyi, M. (1966). *The tacit dimension*. University of Chicago Press.

Schindler, J. (2015). Expertise and tacit knowledge in artistic and design processes: Results of an ethnographic study. *Journal of Research Practice*, *11*(2), Article M6.

Sontag, S. (1966). *Against interpretation: And other essays*. Farrar, Straus & Giroux.

Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199933662.001.0001

Balkan Journal of Philosophy (2026), vol. 18, issue 1, 116 – 136

CAVENDISH AND THE PROBLEM OF INANIMATE MATTER'S MOTION

Miloš Vuletić*

Abstract: This paper discusses a pressing problem that arises within Margaret Cavendish's natural philosophy. Cavendish's conceptions of matter and causation give rise to the problem in question. Inanimate matter is devoid of self-motion and can only be moved by animate matter. Simultaneously, all caused motion is supposed to be occasioned self-motion. Consequently, it appears that the principal conceptions of Cavendish's natural philosophy give rise to an inconsistency. I survey several options for resolving the problem that rely on amendments to Cavendish's conceptions of matter or causation and find them to be unsuccessful. I then propose a solution that makes use of Alison Peterman's interpretation of Cavendish as endorsing compositional motion, and discuss its virtues and implications.

Keywords: Cavendish, materialism, occasional causation, compositional motion.

1. The Problem of inanimate matter's motion

There is a profound problem at the very heart of Margaret Cavendish's natural philosophy. The problem arises once two of the key planks of her position are in place, fully developed and articulated.

The first is Cavendish's particular brand of materialism. According to Cavendish, nature is material through and through; there is nothing in nature that is immaterial or incorporeal.[1] Matter is one in nature, but exhibits compositional variety. In Cavendish's view, there are different "degrees" of matter, differentiated on the grounds of their varying levels of capacity for self-motion. There is *inanimate matter*, utterly devoid of self-motion and characterized as dull, gross, and heavy.[2] And there is *animate matter* which is capable of self-motion. In fact, Cavendish strongly identifies it with self-motion itself.[3] Animate matter appears in two different varieties. One is *sensitive matter*, which Cavendish views, among other qualifications, as being tasked with hauling and moving inanimate matter.[4] The other is *rational matter*, a kind of self-moving matter that moves about unencumbered by inanimate matter.

[1]See, e.g., OEP, pp. 137, 177; PL, pp. 197, 532; GNP, p. 239, etc. I refer to Cavendish's works in the usual way. PPO 2 = *Philosophical and Physical Opinions*, 2nd. ed.; PL = *Philosophical Letters*; OEP = *Observations Upon Experimental Philosophy*; GNP = *Grounds of Natural Philosophy*.

[2]GNP, p. 5.

[3]Expressions of this view are common in Cavendish's corpus. See PPO 2, p. 296; OEP, pp. 205-6; GNP, p. 20. I will use 'degrees of matter' to refer to animate and inanimate matter, unless stated otherwise.

[4]See, e.g., OEP, p. 33; GNP, p. 3.

*University of Belgrade; milos.vuletic@f.bg.ac.rs

The second is Cavendish's view of causation. Although a materialist, Cavendish was no Hobbesian mechanist. It is not the case, in Cavendish's opinion, that because all that is natural is material, then all that exists are the moving and colliding material bodies characterized essentially by their geometrical properties. Her dissatisfaction with the mechanical philosophy ran deep, and her objections were varied and aimed at diverse aspects of it.[5] For present purposes, the most salient aspect of Cavendish's anti-mechanistic outlook is her rejection of the mechanistic notion of causation as transfer of motion. In particular, she had strong objections to the idea that change can be accounted for using causal explanations that invoke "communication by impulse". Her target was the kind of conception that Descartes had expressed in several places. For instance:

> [If] a body collides with a weaker body, it loses a quantity of motion equal to that which it *imparts* to the other body.(AT VIIIA, p. 65, emphasis added)[6]

Cavendish finds this problematic. If motion is a mode of substance, as it is on Descartes' view, then one body cannot impart it to a different body because modes cannot travel from one substance to another. Ultimately, she rejects this type of causation among bodies.

Now, perhaps something other than a body can be an efficient cause of a corporeal effect. Other authors at the time have suggested that there are spiritual agents that do the required work – for instance, something like Henry More's immaterial Spirit of Nature which permeates matter, or perhaps Malebranche's God who is the only true causal agent in the world. For Cavendish, proposals featuring immaterial causes are non-starters. This is not solely a consequence of her materialist outlook. It is entailed by Cavendish's adherence to the view that causes and their effects are alike:

> I cannot conceive how a spirit can have the effects of body, being none it self; for the effects flow from the cause; and as the cause is, so are its effects (PL, p. 197)

The difference in nature between the material and the (supposed) immaterial entities – God included – precludes them from entering into causal relations.[7]

Ultimately, Cavendish finds that she can take only one route once she rejects these alternative options. In the case of body/body causation, there are only material causes and effects involved, but there is no efficient causation in the sense that one body imparts motion to another. Rather, in all such cases it is the self-motion of a body that is the driving force of change. What we perceive as efficient causes of corporeal effects are only [8] for bodies to produce a certain motion themselves:

> When I throw a bowl, or strike a ball with my hand, whether the motion, by which the bowl or the ball is moved, be the hands, or the balls own motion? (...)

[5](James 1999, pp. 222–25) offers a good overview of Cavendish's reasons for her rejection of mechanism.

[6]Compare also (AT XI, p. 15).

[7]Of course, this gives rise to another complaint directed at Descartes: Cartesian mind/body causation is inconceivable, in Cavendish's opinion.

[8]occasions

> To which, I return this short answer: That the motion by which (for example) the bowl is moved, is the bowls own motion, and not the hands that threw it (…) (PL, pp. 444–45)

In another striking example, Cavendish opines that when someone walks across a field of snow, their footsteps do not produce the prints in the snow. Instead, the snow organizes itself into the shape of the footprints (PL, pp. 104–5). What we consider to be a cause in such cases, is merely an occasion, supplied by one material object, for another material object to exercise its own power of self-motion in producing the effect. Bodies cannot *produce* motion in other bodies; at best, they can trigger them to move in a particular way:

> Wherefore one body may occasion another body to move so or so, but not give it any motion, but every body (though occasioned by another, to move in such a way) moves by its own natural motion (PL, p. 100)

In other words, whenever motion is caused, the causal interaction will, at the very least, include an occasional cause and an effect that is due to self-motion of the body caused to move.

The problem I wish to discuss can now be stated. According to Cavendish, inanimate matter is incapable of self-motion. Inanimate matter is *moved* and carried around by the sensitive animate degree of matter. There is also no transfer of motion between bodies, but only occasional causation. The effect of a causal interaction is the only source of motion proceeding from the cause. If so, how is it possible for animate matter to move inanimate matter?

We can frame the problem as a set of jointly incompatible claims, each of which finds ample support in Cavendish's works:

(1) Inanimate matter is devoid of self-motion.

(2) Animate matter causes inanimate matter to move.

(3) All caused motion is occasioned self-motion.

Each pair of claims entails that the third claim is false. If inanimate matter has no self-motion and animate matter causes it to move, then occasional causation does not extend to all cases of caused motion. If inanimate matter is devoid of self-motion and all caused motion is occasioned self-motion, then it cannot be the case that animate matter causes inanimate matter to move. Finally, if animate matter causes inanimate matter to move, and all caused motion is occasioned self-motion, then inanimate matter must have some semblance of self-motion. So it had better be the case that Cavendish did not hold all three claims to be true.

This problem – let's call it the problem of inanimate matter's motion – is significant because it strikes at the heart of Cavendish's conceptions of matter and causation. Cavendish herself was aware of the problem, as we will see, and so are her interpreters. But, remarkably, no detailed treatment of the problem is to be found, let alone a settled-upon solution. This paper aims to rectify this situation and to resolve the apparent inconsistency in Cavendish's natural philosophy.

I will proceed by, first, discussing Cavendish's treatment of the problem, and then by evaluating the prospects of resolving the problem by denying one of the theses (1)–(3).

1.1. Cavendish on the problem

Cavendish provides a clear and succinct formulation of the problem of inanimate matter's motion:

> How it was possible, that that part of matter which had no innate self-motion, could be moved? For (…) if it be moved, it must either be moved by its own motion, or by the motion of the animate part of the matter: By its own motion it cannot move, because it has none: but if it be moved by the motion of the animate, then the animate must of necessity transfer motion into it; that so, being not able to move by an innate motion, it might move by a communicated motion. (OEP, pp. 26–7)

Cavendish puts forth a dilemma. Inanimate matter can move either by its own motion, or by communicated motion. It has no self-motion, says Cavendish, so it cannot move itself. The first horn of the dilemma is thus rejected. What is left is the possibility that its motion could conceivably come from the animate degree of matter by some sort of communication. However, this is not how Cavendish thinks this causal interaction takes place. The above quotation appears in "An Argumental Discourse", a part of Cavendish's *Observations upon Experimental Philosophy*, in which Cavendish stages a dialogue between her metaphysical position (expressed by what she refers to as her "former thoughts") and the other "rational part" of her mind which raises doubts and complaints about this position (referred to as "latter thoughts"). Immediately after raising the problem of inanimate matter's motion, Cavendish's former thoughts respond to the dilemma raised by her latter thoughts:

> The former thoughts answered, that they had resolved this question heretofore, by the example of a horse and a man, where the man was moved and carried along by the horse, without any communication or translation of motion from the horse into the man: Also a stick, said they, carried in a man's hand, goes along with the man, without receiving any motion from his hand. (OEP, p. 27)

The two metaphors are apparently supposed to resolve the problem. Inanimate matter, Cavendish seems to claim, is carried by animate matter much like a rider is carried by a horse, or a stick carried in a hand. There is no transfer of motion involved, so Cavendish rejects the second horn of the dilemma as well as the first. Yet obviously there is some motion in which the rider and the stick are involved: they will both end up at a different place, at least. So how is this supposed to resolve the problem?

It is fair to say that interpreters are not impressed by Cavendish's illustration. Marcy Lascano notes that Cavendish's examples "may not be particularly satisfying" as an account of how the inanimate matter is moved.[9] Jonathan Shaheen proposes

[9](Lascano, 2023, p. 10).

that the metaphors are meant to show that no quantity of motion is lost when the horse carries the rider or the hand carries the stick.[10] Presumably, if there is no loss of motion, then no transfer occurs either. Still, even if successful in showing that no motion was transferred, the metaphors do not discharge Cavendish's task of providing an actual account of inanimate matter's motion.[11]

2. Avoiding the contradiction

2.1. Inanimate matter?

The most obvious and easiest way of avoiding the inconsistent triad is to deny that inanimate matter is completely devoid of self-motion. If that were so, if inanimate matter were to have *some* sort of self-motion, it would be relatively straightforward to explain how sensitive degree of animate matter can move it around – it would amount to another case of one part of matter occasioning another to move on its own accord. At the same time, inanimate matter's self-motion could be so diminished and slow in comparison to animate matter's self-motion that it would not be too difficult to retain the idea that inanimate matter is appropriately qualified as heavy, dull, and gross, as Cavendish is accustomed to do. After all, matter, in her view, has *degrees*, so even the difference between the animate and the inanimate, the most lively and the most sluggish degrees of matter, should still be *a matter of degree*.

This possibility gains some potential support from places where Cavendish mentions degrees of inanimate matter. Perhaps the recognition of degrees of inanimate matter makes room for a continuum of varieties of self-moving matter which extends beyond rational and sensitive degrees. However, this would require an inadvisable overstretching of textual evidence. For example, Cavendish mentions spongy, porous, light (PPO 2, p. 18) and rare (PPO 2, p. 182) degrees of inanimate matter, having in mind gradability with respect to density and rarity, rather than degree of (self-)motion.[12]

In any case, textual evidence for the claim that inanimate matter is devoid of self-motion is overwhelming. Cavendish defines inanimate matter as having no motion of its own:

> ... by Inanimate matter, I understand, that Grosser part of Matter, which is Destitute of all Motion, and as it were Dead, yet apt to receive Life and Motion from the Animate matter that Works upon it. (PPO 2, p. xxxii)

[10](Shaheen, 2019, p. 3557).

[11]Shaheen acknowledges as much when he defers an attempt to solve the problem of inanimate matter's motion to another occasion.

[12]See also PL, 417, where Cavendish emphasizes that the purest degrees of inanimate matter come next to the animate "not in motion, but in the purity of its own degree".

> ...the inanimate part of matter, which in its own nature is moveless or destitute of motion, and is carried along with, and by the animate parts of matter. (OEP, p. 235)[13]

This point is underscored by the fact that Cavendish stresses that the distinction between animate and inanimate matter consists precisely in the former's being self-moving (OEP, p. 156).[14]

It is not just that inanimate matter lacks self-motion by its own nature. This nature does not change in the mixture with the animate degree:

> ...not that the animate matter transfers, infuses, or communicates its own motion to the inanimate; for this is impossible, by reason it cannot part with its own nature, nor alter the nature of inanimate matter, but each retains its own nature; for *the inanimate matter remains inanimate, that is, without self-motion* (PL, p. 99, emphasis added)[15]

The immutability of nature is particularly salient in this context because Cavendish points out that even inanimate matter never rests, due to its being inextricably intertwined with animate matter.[16] This, however, does not mean that in the mixture with animate matter it somehow alters its nature and becomes capable of self-motion.

Denying inanimate matter's intrinsic immobile nature is thus not the route to take in avoiding the inconsistent triad that gives rise to the problem of inanimate matter's motion.

2.2. No causation?

The uncomfortable inconsistency could be avoided if animate and inanimate matter did not engage in causal relations. Then instead of by way of causation, inanimate matter would be moved in some other way, thus preserving the other two theses of the triad. Is there any reason to think Cavendish might have thought so?

One way of arriving at an affirmative answer is contained in the Argumentative Discourse's continuation of the discussion of the problem of inanimate matter's motion. Cavendish's latter thoughts do not accept the invocation of metaphors (of a horse and its rider and of a hand carrying a stick) as acceptable in solving the problem. All the elements of the metaphors are actually examples of "parts or creatures of nature", thus being mixtures of animate and inanimate matter and thereby capable of self-motion. The two metaphors are not adequate insofar as they are representations of a relation obtaining between intrinsically self-moving and intrinsically motionless degrees of matter. Cavendish's former thoughts concede the point and then draw a distinction between two separate notions of a part of nature. Occasional causation holds for a particular kind of parts of nature:

[13] See also PL, pp. 99, 433, 460-1; GNP, p. 21.
[14] Note also OEP, p. 206: "when I speak of self-motion, I name it animate matter."
[15] Compare PL, pp. 515 and 532.
[16] See, e.g., PL, p. 512.

> But, as we told you before, this is to be understood of the parts of the composed body of nature, which, as they are nature's creatures and effects, so they consist also of a commixture of the aforementioned degrees of animate and inanimate matter; but our discourse is now of those parts which do compose the body of nature, and make it what it is: and, as of the former parts none can be said moved, but all are moving, as having self-motion within them; so the inanimate part of matter, considered as it is an ingredient of nature, is no ways moving, but always moved. The former parts being effects of the body of nature, for distinction's sake may be called effective parts; but these, that is, the animate and inanimate, may be called constitutive parts of nature. (OEP, p. 27)

Parts of the "composed body of nature" are such that self-motion belongs to them, and they can enter causal relations on the occasionalist understanding thereof. Examples of parts of nature are, in this case, a horse, a man, a hand, a stick. In each of these parts of nature there is a commixture of animate and inanimate matter. Since there is animate matter as a component of these parts, they can move themselves. But there is also a *different* notion of a part, what Cavendish calls here "constitutive parts of nature", in contrast to "effective parts" of nature (or effects of the body of nature). Constitutive parts are the animate and the inanimate degrees of matter. The distinction at least suggests that it need not be the case that what holds for the effective parts of nature must also hold for the constitutive parts. So perhaps Cavendish had limited the causal relation to the effective parts of nature, denying that the constitutive parts could be considered as its relata. Causal relations, we might say, are at work only at the level of macroscopic entities that are proper parts of nature; what takes place at the constitutive level is not subject to the same laws.

Cavendish's conception of the complete blending of degrees of matter appears to lend some plausibility to this line of thought. In Cavendish's view, the two degrees of matter, animate and inanimate, are so closely mixed

> that no particle in nature can be conceived or imagined, which is not composed of animate matter, as well as of inanimate. (OEP, p. 158)

In this mixture, the degrees of matter are so inseparably intertwined that, in some sense, they could not be separated:

> these degrees or parts of matter, were so closely intermixt in the body of nature, that they could not be separated from each other, but did constitute one body, not only in general, but also in every particular; so that not the least part (if least could be) nay not that which some call atom, was without this commixture; for, wheresoever was inanimate, there was also animate matter (OEP, pp. 34-5)

One way to understand what Cavendish is saying here would be as follows. The infinite matter is homogeneous in the sense that each and every segment of it is such that it contains both animate and inanimate matter. Eileen O'Neill seems to suggest something like this reading:

> the mixture of animate and inanimate matter is not simply juxtaposition or meeting at a surface. For if we took a tiny portion of animate matter which was so

> joined to a tiny portion of inanimate matter, we could still find a tinier subsection of the former that was not in contact with any subsection of the latter. But complete blending requires that any particle you pick, no matter how small, will be composed of both types of matter. (O'Neill 2025, p. 234)

If I understand her correctly, O'Neill thinks that Cavendish's complete blending of the degrees of matter requires that any spatially distinct segment of the material plenum is such that it contains both animate and inanimate matter. "Particle" here would refer to any extended – and therefore divisible – segment of the infinite matter. In case this is what Cavendish has in mind, then we cannot speak of purely animate or purely inanimate parts of matter in the sense of parts that are animate (or inanimate) through and through because each candidate for such 'pure' particle would have to contain *both* degrees of matter. Any talk of purely animate or purely inanimate matter would necessarily refer to an abstraction that cannot be found in nature. From this we could further infer that no causal relations obtain among the degrees of matter. Abstractions do not enter causal relations. This would, in turn, be the reason why Cavendish would invoke the distinction between effective and constitutive parts in the course of her discussion of the problem of inanimate matter's motion: however it is that inanimate matter is moved, it is not by way of causation.

Unfortunately, interpreting Cavendish along these lines is quite unappealing. Let me first address the question of the complete mixture of degrees. The above interpretation of Cavendish's conception of complete blending leads to unacceptable consequences. If each spatially discernible segment of matter is constituted by animate and inanimate matter, then the structure of nature is gunky, in David Lewis's sense of the term.[17] That is, every part of nature has a proper part. And every such proper part is composed of bits of animate and inanimate matter, each of which, in turn, has proper parts also composed of both degrees; and so on ad infinitum. But if this were so, then there would be no way to make any qualitative distinction whatsoever between any two segments of the material plenum. Whichever segment one chooses, it will have the same composition: it will consist of an infinite number of homoeomerous parts composed of both degrees of matter. We could not discern between different parts on the grounds of, say, preponderance of one degree of matter over another. The reason for this is that there cannot be such preponderance given the internal structure of each segment of matter. Suppose we thought that there was, in one region of the plenum, a greater quantity of animate matter than inanimate matter. Perhaps this is how we try to ground the manifest variety we encounter in nature. This would be impossible if matter were gunky in the sense described above and suggested by O'Neill. The supposed higher concentration of animate matter would not be possible given the fact that, as per the target view's understanding of the complete blending, each and every bit of matter, even in the regions of supposed unequal concentration of degrees, has exactly the same gunky structure made up of both degrees of matter. If complete blending is understood as positing that every region of the infinite matter is composed of animate and inanimate matter, Cavendish's Nature would amount to a kind of Anaxagorean primordial mixture in which everything is mixed together

[17](Lewis, 1991).

in such a way that no qualities can be manifest. Such a consequence amounts to a reductio of the interpretation under consideration, given Cavendish's commitment to manifest variety of natural phenomena.

It is thus better to interpret Cavendish as saying that it is *the effective parts* that are always blends of animate and inanimate matter, that we cannot speak of effective parts whose internal structure is devoid of either degree of matter.[18] Effective parts can have constitutive parts that are purely animate or purely inanimate. These constitutive parts can be configured, structured, and concentrated in various ways, thus making possible an explanation of the manifest variety of the material world as a function of effective parts' constitution. They would also be concrete, not abstract. As a matter of fact, constitutive parts so viewed reveal a more promising way of rejecting the second claim of the inconsistent triad. Only those parts of matter that possess animate matter can enter causal relations, because this is what non-mechanical, occasional view of causation requires. Effective parts possess animate matter. This does not hold universally for the constitutive parts – some of them are entirely inanimate, and as such cannot enter causal relations in the world of Cavendish's natural philosophy. The problem of inanimate matter's motion is thus dissolved.

I wish to raise three issues for this way of avoiding the inconsistent triad. For one thing, there is virtually no textual support beyond the Argumental Discourse to be cited in favor of it. The distinction between constitutive and effective parts of nature is virtually nonexistent outside of the exchange between Cavendish's former and latter thoughts. Moreover, and more importantly, Cavendish couches, without exception, the interactions between animate and inanimate matter in causal terms. Animate matter "forces or causes" the inanimate to work with it (PL, 99), it "causeth that part as the inanimate part to help in the Creation" (PPO 2, p. 24). Inanimate matter is "onely the instrument which the Animate works withal" (PL, p. 531).[19] Cavendish describes the sensitive degree of matter, tasked with moving the inanimate matter around, as "the labouring part" (e.g., GNP, p. 21), and its labouring role is its defining characteristic:

> there can be but two sorts of the Self-moving Parts; as, that sort that moves entirely without Burdens, and that sort that moves with the Burdens of those parts that are not Self-moving. (GNP, p. 3)

Addressing the proposed solution more directly, another point can be raised against it. Even if Cavendish did conceive of the relation between constitutive parts as being exempt from causal connection, it remains a complete mystery what should be an alternative way of thinking of the nature of their interaction. In what way could she have possibly imagined their interaction if not on the model of the relation of cause and effect? This is particularly pressing in light of the fact that she employs causal talk when describing the interplay of the animate and the inanimate degrees of matter.

On balance, it is preferable to look for a different way of dissolving the problem than by way of denying that the two degrees of matter are causally connected.

[18] An argument for a similar position is expounded in (Shaheen, 2019), especially pp. 3558–3563.
[19] See also PL, p. 461.

2.3. Is all causation occasional?

Finally, the inconsistent triad might be avoided by rejecting the third thesis, i.e., the claim that all caused motion is occasioned self-motion. So even if animate and inanimate matter do enter into causal relations, perhaps self-motion is not required for inanimate matter to be moved.

One way of rejecting the third thesis is to claim that there are cases of caused motion in which the cause does not merely serve as an occasion for the body that gets put into motion to exercise its own self-motion, but also brings about some sort of change in the body whose motion it instigates. On a view of this sort, bodies in Cavendish's material plenum would not be causally isolated from one another.

It is not particularly controversial that Cavendish acknowledges the existence of causal relations which are not purely occasional. Her objections to body/body causal interactions of the mechanical philosophers are largely concerned with their invoking of transfer of motion as an explanation of how bodies causally affect one another. In Cavendish's view motion and matter are inseparable, so if there is transfer of motion, then there must be transfer of matter as well:

> my opinion is, that if motion doth go out of one body into another, then substance goes too; for motion, and substance or body, as aforementioned, are all one thing, and then all bodies that receive motion from other bodies, must needs increase in their substance and quantity, and those bodies which impart or transfer motion, must decrease as much as they increase… (PL, p. 98)

As Deborah Boyle points out, when Cavendish rejects the transfer of motion, she rejects the claim that bodies can transfer motion alone to other bodies.[20] In certain cases of causation, such as digestion and respiration, some sort of emission of bits of matter does occur as part of the causal process.[21]

Transeunt causation of this kind, if applied to the problem of the inanimate matter's motion, does not achieve much as an explanation of the target phenomenon. Suppose some bits of animate matter do transfer to a portion of inanimate matter. It is not clear how that would help us understand the way in which inanimate matter is set into motion, in addition to not being explanatory of the kind of course inanimate matter might take when moved. Moreover, transfer of motion together with substance cannot explain inanimate matter's motion because Cavendish's former thoughts in the Argumental Discourse use the metaphors of the horse and the rider and of the hand and the stick precisely to show that inanimate matter's motion is not a case of transfer of motion:

> the man was moved and carried along by the horse, *without any communication or translation of motion from the horse into the man* (OEP, p. 27, emphasis added)

[20](Boyle, 2018, p. 100).

[21]See O'Neill's Introduction to OEP, p. xxxv. Cf. also (Lascano, 2023, p. 104) and (James, 1999, pp. 241–42).

The horse and the man represent here, respectively, the animate and the inanimate matter; however it is that Cavendish conceives of their causal relation, it does not involve transfer of motion, regardless of whether it is accompanied by transfer of substance or not.[22]

Another ground for rejecting (3) would be to defend the claim that Cavendish allows for cases of causation in which one body can cause change in another body *without* transfer of matter. A view of this kind was espoused by Eileen O'Neill. According to O'Neill, Cavendish thought that bodies can, in addition to providing occasions for other bodies to move, produce changes on the surfaces of bodies they interact with.[23] So, for instance, a hand can "effect change in local motion" on the surface of a ball it pushes. Perhaps animate matter can do something similar to inanimate matter when putting it into motion.

Cavendish's views on causation have been subject to close scrutiny in recent years. It would take us too far afield to grapple with all the intricacies of the interpretive debates.[24] However, even without such engagement with the topic of causation in Cavendish in general, it is worth mentioning two points that provide grounds for rejecting the current proposal for relieving the tension that claims (1)-(3) give rise to. For one thing, Colin Chamberlain makes a very strong case for the assessment that relevant textual evidence, when fully considered, rules out O'Neill's view.[25] But even if this were not the case, even if there is some room for reading Cavendish along the lines of O'Neill's proposal, this again does not help with the problem of inanimate matter's motion. O'Neill's discussion deals solely with cases of causal connection between *effective parts* of nature. A hand moving a ball, or rolling a cylinder, are instances in which there is, according to O'Neill, causal influence on the surface of the object set in motion, but *also* self-motion of the affected object. What would be needed for this kind of view of causation to provide relief with respect to the problem of inanimate matter's motion, is that Cavendish allows for bits of animate matter to make some superficial changes in bits of inanimate matter and thus produce the latter's motion without transferring any motion. There are simply no explanatory resources in Cavendish's natural philosophy that can make sense of such causal interaction.

None of the ways of avoiding the inconsistent triad that I have discussed so far is successful. Should we then conclude that Cavendish's natural philosophy is saddled with an inconsistency that cannot be removed? I think not. There is yet another way to try to dissolve the problem and I think it is the most promising one. As we shall see, it does come at a price – Cavendish could escape the inconsistent triad but would immediately face certain novel difficulties. But at least the problem in question would be one of committing to an unusual, radical position (which is something that Cavendish would likely not shy away from), rather than facing an outright inconsistency.

[22] See Alison Peterman's discussion of a related point in (Peterman, 2025, p. 109).

[23] See (O'Neill, 2025, p. 243).

[24] Cavendish's account of causation receives extensive treatment in, e.g., (Adams, 2016), (Boyle, 2018), (Lascano, 2021), (Detlefsen, 2006), (Peterman, 2025), and (Chamberlain, 2024).

[25] (Chamberlain, 2024, pp. 106–12).

3. Passive motion and compositional motion

Let us go back to the metaphors by means of which Cavendish attempts to resolve the problem of inanimate matter's motion. We saw that the metaphors, taken at face value, do not offer a straightforward solution to the problem: they are supposed to show how inanimate matter can be moved, yet invoke only effective parts of nature which have self-motion of their own. But they do suggest some key insights into the elements that the solution of the problem of inanimate matter's motion must have. For one thing, inanimate matter is not moved by transfer of motion – Cavendish rules this out explicitly. In addition, Cavendish insists that the inanimate matter is *passive* in its motion; inanimate matter is *moved*, it is *carried* much like a horse carries a rider or a hand carries a stick. So inanimate matter ought to be passive in its motion, and it should be moved without transfer of motion. When looked at from the point of view of Cavendish's opinions on matter and causation, these desiderata seem to paint Cavendish into a corner. Causation without transfer of motion can only be occasional causation, and inanimate matter cannot be caused to move in such fashion, given its lack of self-motion. So it seems we have arrived at the same impasse yet again.

But even though I have framed the problem as arising from Cavendish's conceptions of matter and causation, its solution need not be directly concerned with these conceptions. We can also frame it as a problem concerning motion. *What kind of motion* can we ascribe to inanimate matter anyway? Once we ask ourselves this question, I think we can finally move in the direction of a promising solution.

3.1. Compositional motion

Cavendish does not discuss the nature of motion as explicitly and systematically as do some of her contemporaries – Hobbes and Descartes, for instance. Still, many important commitments regarding motion can be recovered from her works. In an important recent contribution, Alison Peterman has called attention to Cavendish's conception of motion as *compositional*.[26]

Hobbes, one of the authors Cavendish extensively engages with, treats local motion as change of place. In order to do so, he distinguishes between the real space of a body and the place of a body. The real space is the volume a body takes up. This is just the three-dimensional extension of a body, or its magnitude. On the other hand, place is "that space …which is coincident with the magnitude of any body". Place is "a phantasm" of any body, while real space/magnitude is an accident of every body. Also, place is nothing out of the mind, while real space is "nothing within it" (DC 8.5, p. 106). Cavendish denies the real space/place distinction. Criticizing Hobbes, she expresses her view that place and magnitude should be identified:

> I believe that Place, Magnitude and Body are but one thing, and that Place is as true an extension as Magnitude, and not a feigned one; Neither am I of his opinion, that Place is Immoveable, but that place moves, according as the body moveth, for not any body wants place, because place and body is but one thing,

[26] See (Peterman, 2019) and (Peterman, 2025, pp. 93–101).

> and wheresoever is body, there is also place, and wheresoever is place, there is body, as being one and the same (PL, p. 56)

Noticing that Cavendish rejects the real space/place distinction, Peterman argues that Cavendish conceives of actual motion as *compositional motion*. For Cavendish, on this interpretation, facts about motion are grounded in facts about division and composition. This is corroborated by numerous instances in which Cavendish appears to endorse such a view; for example:

> a man sailing in a ship, cannot be said to keep place, and to change his place; for it is not place he changes, but only the adjoining parts, as leaving some, and joining to others; and it is very improper, to attribute that to place which belongs to parts, and to make a change of place out of change of parts. (PL, p. 106)

> a man goes a hundred miles, he leaves or quits those parts from whence he removed first; but as soon as he removes from such parts, he joins to other parts, were his motion no more than a hairsbreadth; so that all this journey is nothing else but a division and composition of parts, wheresoever he goes (OEP, p. 127)

According to Peterman, in texts like these Cavendish states the view according to which motion is reduced to changes in mereological facts. In these passages Cavendish writes not as though division and composition necessarily accompany motion, but implies, in Peterman's view, that something's motion is grounded in (or just *is*) its quitting some parts and joining others.[27] Peterman accordingly characterizes Cavendish's compositional motion thus:

> For a body A to move is for it to go from being a part of some larger body, X, to being part of a different larger body, Y. (Peterman, 2019, p. 483)

In addition to this, Cavendish repeatedly stresses that an actual bit of matter's dividing from one body and joining another is a single act. For instance:

> there is as much composition, as there is division in nature; and as soon as parts are divided from such or such parts, at that instant of time, and by the same act of division, they are joined to other parts, and all this, because nature is a body of a continued infiniteness, without any holes or vacuities. (OEP, p. 127)

In fact, in some places Cavendish goes as far as to reduce all change in nature to changes in parthood relations, and elsewhere says that the chief actions of nature are to divide and to unite, and lists this as one of the principles of her philosophy.[28]

Further evidence – albeit indirect – for Cavendish's espousing of compositional motion is supplied by the fact that she was not alone in doing so. Daniel Whiting has shown convincingly that Kenelm Digby was also a proponent of a view of this kind.[29] As a matter of fact, Digby's proposal predates Cavendish's, and it is likely

[27] "[Cavendish] implies that something's motion just is, or just is grounded in, its quitting some parts and joining others" (Peterman, 2019, p. 483).

[28] For the former claim, see (OEP, pp. 78, 238; GNP, p. 27), and for the latter (OEP, p. 190).

[29] See (Whiting, 2024).

that he influenced her.[30] Similarly to Peterman's interpretation of Cavendish, Whiting concludes that Digby "takes local motion not only to correlate with change in parthood, but to consist in it".[31]

Local motion that consists in division and composition requires a very low bar to be set regarding the conditions of unity of objects. In Digby, this is explicitly present in his claim that bodies unite to form one body when they make contact:

> [I]n fact, [Digby] makes a stronger claim that, when two bodies make contact, they unite to form one body. For 'any two partes that are indistant from one an other', Digby writes, 'theire owne nature maketh them one' (…) More straightforwardly: 'Two bodies can not touch one an other, without becoming one'. (Whiting, 2024, p. 7)

The joining of two bodies into a new unified whole requires only that they touch each other. It is plausible that Cavendish would adopt a similar position given her conception of division and composition as universally concomitant. If mutual separation of two parts *by the same act of division* unites them with other parts, then it appears that contiguity is sufficient for unity – perhaps barring the possibility that bodies in question are somehow incapable of unifying, say because of mutual antipathy.[32]

3.2. Compositional motion and local motion

In case local motion always entails changes in parthood relations and vice versa, we can begin to see how inanimate matter could be moved. If all motion is grounded in, or identical with, facts about parthood relations, division and composition, then we can think of the difference between animate and inanimate matter with respect to motion in the following way. Animate matter, as an active principle of Nature, is that constitutive part that engages in divisions and compositions of parts of nature that we recognize as motion and change. Inanimate matter, as a passive principle of Nature, is what is involved in divisions and compositions, but this being so is *not* a matter of its having any influence on the actions of division and composition; it is merely subject to such changes.[33]

On this view, being moved would then amount to being passively involved in division and accompanying composition of parts of nature. Divisions and compositions are directed by the rational part and carried through by the self-moving sensitive matter which "lugs" inanimate matter around in the sense that it makes certain portions

[30] On this point, see (Whiting, 2024, pp. 14–15).

[31] (Whiting, 2024, p. 8).

[32] On the use and role of sympathy and antipathy in Cavendish, see (James, 1999, pp. 237–42) and (Boyle, 2018, pp. 102–4).

[33] Peterman notes that, for Cavendish, self-motion is "the power that a bit of matter has to set itself in motion," and that this allows for a distinction between self-motion as motive force and compositional motion. Furthermore, the distinction in question "goes some way toward explaining how inanimate matter can be moved without self-motion" (Peterman, 2019, pp. 489–90). I take my proposal in sections 4.2 and 4.3 to be a development of Peterman's suggestion, even though I do not, as will become clear, share all of Peterman's views on motion in Cavendish.

of inanimate matter constitutive parts of new parts of nature. When a man moves, he leaves certain parts of nature and joins others. By becoming constitutive parts of the newly formed effective parts of nature, portions of inanimate matter are "moved" without there being any transfer of motion: it is all division and accompanying composition.

But more needs to be said about how exactly this is supposed to work. And here I think we should be careful when ascribing overly strong claims to Cavendish (and Digby, for that matter). I think Peterman is right that for Cavendish local motion entails compositional motion. When Margaret travels from London to Antwerp, she does, in Cavendish's view, quit "those parts from whence [s]he removed first" and join other parts. And conversely, when there is joining or parting of bits of matter, there is motion in the sense that the place of an object is altered due to its increasing or diminishing, given the idea that an object's place is inextricable from its magnitude. But we should refrain from asserting that for Cavendish local motion of a body *just is* (or is grounded in) "its quitting some parts and joining others".[34] The reason for this is that division and composition alone cannot explain how Margaret ends up in Antwerp. In order for her to undergo a change from being a part of some larger body X in London to being a part of some larger body Y in Antwerp, she must relocate, i.e., she must somehow *move* from the vicinity of some particular bodies she touches and is united with, and arrive in the vicinity of other bodies to be united with. This requires a change of location, even if we cannot – according to Cavendish – talk about change of place. There must be some semblance of translational motion if some objects in London are to be left behind and different objects on the way to Antwerp joined. Peterman surely uncovers something important when she points out that Cavendish is opposed to motion as change of place, and that Cavendish invites us to think of motion in terms of division and composition. But this need not entail that local motion is *nothing but* division and composition. It would be very bad if Cavendish thought that division and composition were sufficient for motion of the kind that gets one from London to Antwerp because division and composition cannot amount to nor produce translational motion.

Luckily, there is no lack of textual evidence in Cavendish's works that opens up the possibility that she was committed to something weaker than outright identification of local motion with division and composition. In the 1663 edition of *Philosophical and Physical Opinions* she plainly talks about local motion as change of place:

> as for Local Motion, it is inherent in all Animate matter, as a Self-moving matter, so that Local Motion is to remove from place to place by Self-motion (PPO 2, p. 38)

Elsewhere, she expresses the view that local motion involves movement to a place, where place is understood in terms of surrounding bodies:

> When I say, an Animal or any thing else that has exterior local motion, goeth or moveth to such or such a place, I mean, to such and such a body; and when

[34](Peterman, 2019, p. 483).

> such a Creature doth not move out of its place, I mean, it doth not remove its body from such or such parts adjoining it. (PL, p. 536)

This is highly reminiscent of Descartes' conception of motion and 'external' place, according to which motion is transfer "of one piece of matter, or one body, from the vicinity of the other bodies which are in immediate contact with it, and which are regarded as being at rest, to the vicinity of other bodies" (AT VIIIA, p. 53). Similarly, Digby talks about a body's location in terms of surfaces of bodies that immediately surround it; in his colorful turn of phrase, these surfaces, and thereby the body's location, are its 'cloathing'.[35]

The case for an interpretation of Cavendish as holding a view of local motion which *does not* reduce it to composition and division does, therefore, have some hope. Granted, a lot more would need to be said in order to rule out Peterman's stronger reading of Cavendish, and I cannot do so here. But one should not discount the fact that it is very difficult – if not downright impossible – to see how compositional motion alone can account for translational motion. I think this alone should grant an exploration of alternative interpretations of Cavendish on motion, especially if those interpretations preserve an important role for division and composition and provide a way of explaining inanimate matter's motion. I will now turn to this final issue.

3.3. Compositional motion as passive motion

Once we allow that Cavendish recognizes compositional motion, we can make sense of inanimate matter's being moved. If it is sufficient for a body to move by being divided from some bodies – and, concurrently, joined with other bodies – then it can move even if it does not engage in any sort of translational motion. It is sufficient that other bodies "leave or quit" it, as Cavendish puts it. Animate matter, presumably, can do so, given that it has self-motion and can therefore initiate division (and composition) by quitting some bodies (and joining others). However, as pointed out above, I do not see how this can ever take place unless there is *some* translational motion, some change of location in addition to division and composition. We have to allow, on pain of leaving Cavendish without a workable conception of local motion, that there is translational motion that accounts for how one object ever removes itself from the vicinity of some objects and arrives in the vicinity of others.

This gives us a pleasing way of differentiating between animate and inanimate degrees of matter and their respective motions. Animate matter is self-moved and thus can initiate motion that propels it from one location to another. As it does so, it engages in divisions and compositions, thus moving compositionally as well. Compositional motion, in turn, is how animate matter can move inanimate matter which is completely devoid of self-motion *and* incapable of altering own nature: when animate matter travels from one location to another, it separates itself from some bits of inanimate matter – among other bodies – and joins with other bits of inanimate matter. Thereby these portions of inanimate matter *move* by virtue of the fact that they stop being parts of one larger body and become parts of a different larger body. This is what *being*

[35] See (Whiting, 2024, p. 5).

moved amounts to. It is passive motion without any transfer of motion – precisely what Cavendish's metaphors in the Argumentative Discourse are supposed to illustrate.

This is what I take to be the best way to avoid the inconsistent triad. Instead of denying (3) by intervening on Cavendish's conception of causation, we can deny it by recognizing that there is caused motion that is *not self-motion*: namely, passive compositional motion.[36]

The solution of the problem of inanimate matter's motion that I propose has some significant virtues. First and foremost, it gives a clear sense, while staying grounded in Cavendish's text, to the idea that inanimate matter can be moved, while upholding Cavendish's accounts of causation and inanimate matter's nature. In addition, it further fleshes out how animate and inanimate degrees of matter differ with respect to their characteristic motions. Animate matter is capable both of propelling itself forward in the sense of having translational motion, and in the sense of having compositional motion; inanimate matter is only capable of passive compositional motion.

But as I indicated in the introduction, it comes at a price. Namely, it requires some striking claims to be ascribed to Cavendish with respect to effective parts' numerical identity over time. Recall the example of Margaret traveling from London to Antwerp. If I am right that Cavendish ascribes no motion other than passive compositional motion to inanimate matter, then those portions of inanimate matter that were in Margaret's vicinity while she was in London are *still over there* even when she finds herself in Antwerp. As a matter of fact, so must be those bits of inanimate matter that partially composed Margaret herself at the time when she was in London. So how is it that Margaret, an effective part of nature composed of both animate and inanimate matter, is now somewhere where *her own* earlier inanimate matter is not? The only way out of this conundrum I can envision is this. What happens is that Margaret's animate matter moves from London to Antwerp, and in the course of doing so it serially joins with and divides from different inanimate parts. The inanimate parts, in the course of this change, move passively insofar as they change parthood relations due to Margaret's animate matter's motion. We can imagine something like a swarm of highly agile and quick bits of animate matter engulfing passive and immobile bits of inanimate matter, animating them by making them temporarily parts of Margaret's body, and then moving on to other bits of inanimate matter while separating from the former ones. The newly encountered bits of inanimate matter now become components of Margaret's body, replacing the ones that Margaret's animate matter has just separated from.

This would require at least two contentious claims to be true. One is that different parts of matter can become parts of a unified body merely by touching. The other is that we can sensibly talk about numerically identical bodies despite the fact that a significant portion of their composition changes with the slightest of movements – i.e., the inanimate matter that moves compositionally when animate matter initiates its own motion.

[36] It is worth noting that Cavendish describes inanimate matter as "a dull, *passive and moved* degree of matter" (OEP, p. 30, emphasis added).

The first claim seems to be implied by the very notion of compositional motion. As Cavendish says in her example of a traveling man: "as soon as he removes from such parts, he joins to other parts, were his motion no more than a hairsbreadth" (OEP, p. 127). Compositional motion so described requires only quite weak conditions to be placed on a body's unity. But the second claim is perhaps even more controversial. Is there any reason to think that Cavendish may have subscribed to it, other than its being implied by a promising solution to an apparent inconsistency?

Here it is difficult to be overly confident that textual support is forthcoming. Even in the absence of the decidedly unorthodox view of an object's identity over time implied by my proposal, Cavendish's views on individuation of material objects is difficult to pin down accurately.[37] I do wish to highlight an interpretive direction in which one could find textual support for the view I am proposing.

In her mature works of the 1660s, Cavendish repeatedly uses the notion of "corporeal figurative motion". It plays different explanatory roles, one of which is to help explain perception and knowledge. But it also serves to explain variety in nature and individuation of objects. For instance, Cavendish equates corporeal figurative motions with Nature's finite parts:

> what we call ***finite parts, are nothing else but several corporeal figurative motions***, which make all the difference that is between the figures or parts of nature, both in their kinds, sorts, and particulars. (OEP, p. 31; emphasis added)

But she also identifies corporeal figurative motion with the animate degree of matter:

> wheresoever was inanimate, there was also animate matter; which animate matter, was nothing else but corporeal self-motion ... self-moving matter, or corporeal figurative self-motion (OEP, pp. 34–5)

> the infinite varieties are made by the Self-moving parts of Nature, which are the Corporeal Figurative Motions of Nature (GNP, p. 13)

Elsewhere, the relation between corporeal figurative motion and parts of matter is somewhat altered, so that parts are the *effects* of corporeal figurative motions:

> nature is a perpetually self-moving body, dividing, composing, changing, forming and transforming her parts by self-corporeal figurative motions (OEP, p. 85)

> But if anyone desires to know, or guess at the parts of nature, he cannot do it better, than by considering the corporeal figurative motions or actions of nature: for, what we name parts, are nothing but the effects of those figurative motions; so that motions, figures and parts, are but one thing. (OEP, p. 162)

So corporeal figurative motions are identified with animate matter, and are also either identified with effective parts of matter, or with their causes. In either case, it seems that effective parts of nature, even though they are mixtures of animate and

[37] Colin Chamberlain is quite right to describe individuation of bodies in Cavendish's system as "messy" (Chamberlain, 2024, p. 120).

inanimate matter, are such that their identity is strongly related to the animate degree of matter alone:

> Composed Figures which we name Creatures, are produced by particular Associations of Self-moving Parts (GNP, p. 27)

> [Animate matter] produces all things, creates all things, dissolves all things, and transforms all things in Nature; but not out of Nothing, or into Nothing, as to create new Creatures which were not before in Nature, or to annihilate Creatures, and to reduce them to nothing; but it creates and transforms out of, and in the same Matter which has been from all Eternity (PL, p. 350)

All of this lends some credibility to the idea that the numerical identity of the effective parts of nature could be tied to their animate components, and that inanimate components are inessential (and so, perhaps, subject to substitution with other mutually interchangeable bits of inanimate matter). Of course, this is not a fully worked out account of individuation that would erase doubts about my proposal. But I hope it shows that the proposal is far from being alien to Cavendish's thought as it is expressed in her works.[38]

The proposal I have sketched also helps make sense of certain claims Cavendish makes about motion and inanimate matter. For instance, she claims that "Nature and her creatures know of no rest, but are in a perpetual motion" (PL, p. 447), and in particular that inanimate matter *is never at rest*:

> Your fifth question is, Whether this inanimate Matter do never rest? I answer; It doth not: for the self-moving matter being restless in its own nature, and so closely united and commixed with the inanimate, as they do make but one body, will never suffer it to rest; so that there is no part in Nature but is moving; the animate matter in it self, or its own nature, the inanimate by the help or means of the animate. (PL, p. 512)[39]

If inanimate matter moves compositionally, and does so as a consequence of animate matter's self-motion, then it will quite literally never be at rest. Even the smallest of translational motions of the smallest of portions of animate matter would recompose other material parts – inanimate matter included – and thus move them compositionally. Finally, it becomes clear why Cavendish can say that "it is impossible that the animate part of matter should move without the inanimate" (PL, p. 461): since animate and inanimate degrees of matter are thoroughly mixed within effective parts of nature, every motion of the self-moving degree will entail compositional motion of the neighboring inanimate matter.

[38] One pressing problem that I acknowledge but cannot tackle here is this: how is this sort of movement of swarms of animate matter supposed to work in a material plenum? This is a difficult question, but I think it is a difficult question *for Cavendish*, as it was for Descartes and Hobbes, and that any interpretation of her views on motion must find it troublesome.

[39] See also PL, p. 145.

4. Conclusion

The problem of inanimate matter's motion is a deep and serious one for Cavendish's natural philosophy. It strikes at the very center of her outlook and brings with itself a tension that cannot be relieved without a substantive intervention on one of her key explanatory concepts. I have surveyed the interpretive options for avoiding it, and found that dissolving the problem by revising Cavendish's notion of inanimate matter, or her conception of causation, is doomed to failure. A solution that shows promise, I have argued, is the one that leans on a revision of Cavendish's conception of motion. Acknowledging compositional motion opens the way to a superior solution of the problem. It comes, I believe, at a significant cost: Cavendish would have to be saddled with some prima facie unpalatable and burdensome commitments. I take it that that's the price of consistency of her natural philosophy. A full exploration of the true cost of this solution is a matter for a different occasion. For now it will suffice that the options for resolving the problem, as well as their respective advantages and stakes, are clarified and brought forth.

Acknowledgments:

An earlier version of this paper was presented at the "Understanding vs. Knowledge in Science, Fiction, and Art" workshop at the Sofia University "St. Kliment Ohridski". I thank the audience for discussion. My research has been supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (451-03-33/2026-03/200163).

References

Adams, M. (2016). Visual perception as patterning: Cavendish against Hobbes on sensation. *History of Philosophy Quarterly, 33*(3), 193–214.

Boyle, D. (2018). *The well-ordered universe: The philosophy of Margaret Cavendish*. Oxford University Press.

Chamberlain, C. (2024). Move your body! Margaret Cavendish on self-motion. In S. Bender & D. Perler (Eds.), *Powers and abilities in early modern philosophy*. Routledge.

Detlefsen, K. (2006). Atomism, monism, and causation in the natural philosophy of Margaret Cavendish. *Oxford Studies in Early Modern Philosophy, 3*, 199–240.

James, S. (1999). The philosophical innovations of Margaret Cavendish. *British Journal for the History of Philosophy, 7*(2), 219–244. https://doi.org/10.1080/09608789908571026

Lascano, M. (2021). Cavendish and Hobbes on causation. In M. P. Adams (Ed.), *A companion to Hobbes*. Wiley-Blackwell.

Lascano, M. P. (2023). *The metaphysics of Margaret Cavendish and Anne Conway: Monism, vitalism, and self-motion*. Oxford University Press.

Lewis, D. K. (1991). *Parts of classes*. B. Blackwell.

O'Neill, E. (2025). Margaret Cavendish, Stoic antecedent causes, and early modern occasional causes. In *Disappearing ink: Essays in early modern philosophy*. Oxford University Press.

Peterman, A. (2019). Margaret Cavendish on motion and mereology. *Journal of the History of Philosophy, 57*(3), 471–499. https://doi.org/10.1353/hph.2019.0055

Peterman, A. (2025). *Cavendish*. Routledge.

Shaheen, J. L. (2019). Part of nature and division in Margaret Cavendish's materialism. *Synthese, 196*(9), 3551–3575. https://doi.org/10.1007/s11229-017-1326-y

Whiting, D. (2024). Kenelm Digby (and Margaret Cavendish) on motion. *Journal of Modern Philosophy, 6*(1), 1–27.

Balkan Journal of Philosophy (2026), vol. 18, issue 1, 137 – 146

SCHELLING'S THEORY OF ART AND UNDERSTANDING IN JUXTAPOSITION TO CLASSICAL PLATONISM: A CASE OF CONFLICT BETWEEN PARADIGMS

Nikifor Avramov*

Abstract: The text inquires about a lack in Schelling's thought – a lack that appears as a result of his peculiar use of metaphysical jargon in an otherwise characteristically modern thought. In particular, it reasons through his understanding of the connection between art and understanding in order to illuminate the issues with his approach to ontological ground (*Grund*). The inquiry juxtaposes Schelling's views to those of the classical metaphysics of antiquity in the context of the theurgic forms within the Platonic tradition.

Keywords: art, understanding, symbol, Schelling, metaphysics, Platonism.

1. Introduction

With the present text, I aim to consider a simple hypothesis regarding an aspect of the problematic relationship between pre-modern metaphysics and F. W. J. Schelling's case of a modern attempt at metaphysics. My aim isn't to exhaustively present the two positions but to suggest a possible direction for inquiry when it comes to the further consideration of their relation. I take Schelling's well-known idea of art as "*the organon of philosophy*" (laid out in his early work, "*System of Transcendental Idealism*" (1800)) as a possible example in the juxtaposition. The counterpoint element referred to as "*classical metaphysics*" (a term used by researchers such as W. Newell and A. Hassanpoor, particularly following M. Heidegger) is exemplified here specifically through its Platonic motive of theurgic and symbolic connection with the transcendent (the Hellenic deities). The Platonic engagement of spirituality – as a relation with the transcendent of both theoretical and practical nature – has been studied thoroughly by authors like A. Uzdavynis (e.g., in "*Philosophy and Theurgy in Late Antiquity*") and G. Shaw (whose work focuses on theurgy, e.g., in "*Theurgy and the Soul: The Neoplatonism of Iamblichus*"); not to mention spiritual motives and practices have been a core interest for Platonic thought ever since Plato himself.

Considering the limits of my study, I'll use the general framework of Platonism as a motive without further elaborating on the intricacies of that framework. However, my focus will be primarily on Schelling, as his position is the one I want to problematize against the background of Platonism; and even more, my focus will be on the juxtaposition itself. Therefore, what I mean to shed light on through the juxtaposition

*Institute of Philosophy and Sociology, Bulgarian Academy of Sciences; nikifor.avramov@ips.bas.bg

is the unfolding of a common theme, found in the two otherwise fundamentally different paradigms: one pre-modern, metaphysical, and spiritual in a traditional sense of the word, the other characteristically modern, post-Kantian, and with its own peculiar understanding of the term "*Geist*". My inquiry hypothesizes that this common theme is a perspective from which one can look at structures in the ontology of both art and understanding in relation to each other, through the relationship of each with what could be more impartially called a "*grounding beyond*," grasped by both in their own manner. The continuity between Schelling and Platonism has been repeatedly pointed out in light of a broader intellectual constellation (e.g., when discussing Schelling and certain other late modern thinkers, a researcher has noted that "*their philosophical principles are akin to Neoplatonism's [...] the latter (according to Lossev) is a synthetic expression of antiquity*" (Penev, 2013, p. 31)). Here I aim to explore more so the discontinuity between Schelling and the Platonic tradition in question – to emphasize their difference, though to also present them as a useful pair in the theorizing of art and understanding. Schelling's potentially positive role regarding questions pertaining to the broad tradition of Western spirituality and metaphysics has also been explored by late thinkers, such as E. Voegelin[1]; here I mean to suggest Schelling's negative role regarding this tradition, through a consideration of possible lacks in his own attempt at such.

Furthermore, it's always been somewhat obvious that, prior to modernity, the history of art comes along with the history of religion. The relation between religion and art could be explored further in ontological terms. My inquiry inherently emphasizes another relation as well – the one between the history of thought and that of spiritual traditions. I rely on the general motives of such relations to approach the bridging of the gap between understanding and art in the context of Schelling and the Platonic tradition. A starting point of my inquiry is that both of the paradigms of my interest are oriented toward the consideration of an ontological ground (*Grund* for Schelling, the heights of the metaphysical hierarchies – *to Hen*, etc. – for the Platonists). The grounding relates to the peculiar forms of knowledge and/or practical and existential approaches to that ground presented by each of the paradigms. In particular, through the idea of grounding, Schelling "*generates a new notion of theoretical grounding," pertaining to "relations within a system, as based on basic conceptual divisions striving toward unity*" (Fisher, 2024) – that is, as a theory differing from classical Platonism in positing its fundamental principle of unity. A common intuition and motive of both paradigms is that this ground persists and is present in the unfolding of knowledge, which mirrors it under some kind of principle of correspondence (*adaequatio*). Where their perspectives differ is in how the Platonic tradition – particularly in the case of middle and late Platonism, and even more so in light of it developing a theurgic approach[2] – connects religion/spirituality and philosophy somehow naturally, without forcing the relation (to be precise, it both receives the seeds of this connection

[1] As S. McGuire notes, Voegelin "*puts Schelling in the company of Plato, Augustine, and Aquinas, grouping him with these "spiritual realists," who attempted to balance the tensional forces within their civilizations*" (McGuire, 2012).

[2] Being defined through Iamblichus and late Platonists as a ritualized co-activity with the gods, in a context akin to the well-known Christian liturgy.

historically, from the relations between, e.g., the legacy of the Greek cults of antiquity and Platonic philosophy, and participates in its construction). It's important to note this tradition keeps the two aspects deeply connected (either in subject or in the grounding of the subject), while positing them as characteristically distinct (in method). For Schelling, in turn, the classical forms of spirituality and metaphysics aren't of literal interest in correspondence to his theoretical aims, which are outside or against classical metaphysics, those aims being posited in through a post-critical philosophical project. This shift away from the explicitly spiritual/metaphysical vertical relation with the transcendent defines Schelling's project as being ultimately focused more so on an immanent dynamic ontology. This is also an inherently anthropologically oriented focus – as thinkers like L. Feuerbach will explicate through the same intellectual lineage not much later in history. This focus would explain why, when it comes to Schelling, one doesn't encounter a clear positive contemporary concept of spirituality as a distinct and autonomous cultural field, corresponding to forms of experience irreducible to paradigms outside of spirituality (psychology, anthropology, etc.). Though Schelling does take the Christian church to be an important historical event and a relevant object of study or inspiration for his thought[3], what fundamentally characterizes his thought puts the same outside of the characteristically religious. On a positive note, it's in Schelling that we find a very explicit connection between a peculiar philosophy imbued with spiritual motives and art.

Most importantly as far as my inquiry is concerned, this mixture of topics seems to naturally present a common theoretical space, which is unified by common themes and problems (e.g., and in particular, the problem of grounding, itself tied to other motives of structural importance for metaphysics, spirituality, and art, such as "*depth*" and "*inner*" among others). This implies an underlying "*unified field*" of perspectives on the possible connections between art and understanding. In turn, it also justifies the juxtaposition of the modern Schelling with the classical Platonic tradition of antiquity.

2. The fundamental role of symbol and symbolic consciousness

Just as Schelling seems to do, I conflate art and symbol. Whether all art is symbolic could be questioned; however, I accept this view tactically and consider it adequate even beyond its strict explication by Schelling. E.g., art can be posited as symbolic in order to separate it from pure aesthetics (the latter pertains exclusively to the sensual, whereas art and symbol engage more explicitly in the formation of meaning as well). Moreover, symbolic consciousness could be seen as a fundamental form of consciousness at work in the psychic aspect of the art sphere (as opposed to its material aspect, more so implied by the classical interpretation of the dyad "*techne-poiesis*" throughout antiquity). Defining art through symbol also makes Schelling's philosophy of art relatable to the role of symbol within the Platonic tradition (as *sumbolon*, also sometimes *sunthema*) – which is a topic especially for the middle Platonism's theurgic text "*Chaldean Oracles*" and late Platonism's theurgic approaches of Iamblichus

[3] Particularly in "*The Ages of the World*" and "*Philosophy of Revelation*"

and Proclus. The commonality in question follows the idea that symbol engages, as thinkers like Voegelin have pointed out, a primal and fundamental modality of consciousness: "*there were no ideas unless there were symbols of immediate experiences*" (Voegelin, 2011, p. 95). The aforementioned is why symbolic consciousness is considered as a kind of ground through which philosophical explication builds itself (because philosophy follows what is immediately grasped by means other than those of philosophy). As already suggested, an axis of my inquiry is an intuition implicit in both paradigms (Schelling's and the one of classical Platonism): that symbolic consciousness, being a primal way of materializing the liminal contact of individual consciousness with a beyond into a form (a form readily accessible and tangible for the surface levels of consciousness and for the collective), also bridges the gap between art and understanding. In other words, what's in question in the case of both art and understanding – as long as we think of them in terms of fundamental grounding – would be their standing in tension to symbolic consciousness, itself grasped as a *metaxy* (following Voegelin's term), a relational "*in between*" at the edge of human experience. In such a liminal experience, the human form touches on what is experienced to be an extreme and irreducible otherness beyond common experience. And in a certain sense this beyond – the transcendent – also acts as grounding, framing, or transcendental in a broad sense of the term, which defines the importance of the transcendent for man. Just as the idea of the transcendental is an idea of productivity (in the sense of the production of experience), for Schelling art and symbol are seen also in terms of productivity – something that goes beyond classical conceptualizations of the *metaxy*.

3. Schelling's position

It's well known by researchers that for Schelling, "*art pertains to how humans comprehend themselves and their world*" (Croci, 2024). The core element in Schelling's claim that art is the "*organon of philosophy*" seems to be "*an inaccessible to discursive knowledge ground for such – the intellectual intuition*," which "*demands a reinterpretation of the grasping of intellectual intuition and reflexion in Schelling as two mutually exclusive modes of understanding*" (Kassabov, 2008, p. 52).[4] Regarding art, Schelling himself says that the postulated experience which philosophy requires "*is none other than the product of genius, or, since genius is possible only in the arts, the product of art*" (Schelling, 1978, p. 352). Schelling sees art in light of expressivity, as rooted in a liminal contact with the beyond – therefore as a particular framing of the relative presence of the beyond in terms of manifestation (relative to the relativity of material existence, of manifestation, in relation to the ground). In terms of surface appearances, such a view on art seems divorced from what the opinion of artists in the fields of the arts is, particularly in the modern fields structured around the

[4] This interpretation also emphasizes the relation between art and understanding isn't necessarily linear but ontological in a broader sense: "*once art has played its role in the transcendental defining of the structure of the identity and its actuality, philosophy becomes absolute and can now begin from such structure*" (Kassabov, 2008, p. 61).

emancipation from metaphysics. Yet Schelling's view seems to be penetrating to the intimate core of art in such a manner that it illuminates something important about it – by relating art to a depth that, while being intimately its own, also transcends it (as, to use Jacques Lacan's term, "*extimate*"). If one wants to understand art in terms of grounding meaning, such a formula seems crucial. In the context of art, the formula of Schelling's relation between what he calls by different names in different texts, such as, for example, "*conditional*" and "*Unconditional*" or "*Absolute*" is as follows: "*The artistic symbol [...] doesn't refer to the general but is the general. [...] it's an idea and a full – even if as a particular – identity between general and particular. That it's a full identity also means that it's identical to the Absolute*" (Kassabov, 2008, p. 66).

Later in time, in his "*Philosophy of art*" (1802–3, 1859 in publishing), Schelling also explicates the symbolic basis of art. The transition between the two texts isn't my topic here; it suffices to note that the early Schelling thought that "*only art can reveal the Absolute*" (Bowie, 1993, p. 76), while the later Schelling saw art as a possible form of such realization. Either way, by referring to ancient Greek mythology in the "*Philosophy of art*", he transposes the "*allegorical*" and the "*schematic*" in "*The absolute indifference of both: the symbolic. [...] Homer did not first render these myths independently poetic and symbolic; they were such from the very beginning*" (Schelling, 1989, pp. 47–48).

Correspondingly, in the context of grounding, art is a sort of creative extraction accessible to the genius; which is why "*The separation of the allegorical element in them [...] occurred to a later period, something possible only after all poetic spirit was extinguished*" (ibid.: 48).

4. Schelling's lack

The variety of religious/spiritual motives in Schelling (not just his use of terms like "*Absolute*" but also his aforementioned interpretation of mythology, among other motives) isn't there by chance but both by historical reception and by theoretical role in his own attempt at metaphysics. In the context of the attempts, the motives suggest a specific function that needs to be addressed – a function that has previously occupied a domain other than that of pure philosophy (which, as Voegelin at times remarks when speaking of modern intellectual Gnosticism[5], the German idealists seem to pursue as a gnoseo-sophy and *Wissenschaft*). Moreover, certain motives from past spiritual traditions could address the structuring of art in its role as beholder of the tension with the Absolute – however without explicating this theoretically as a theory of creativity and bringing forth (as it is for Schelling, that is, also in the context of art as it's defined in a characteristically modern context). The structure of the bringing forth is one reason why both Schelling himself (since his earliest texts) and Schelling's reception,

[5]E.g., by identifying the immanentist tendency as gnostic, for Voegelin "*gnosis may be primarily intellectual and assume the form of speculative penetration of the mystery of creation and existence, as, for instance, in the contemplative gnosis of Hegel or Schelling*" (Voegelin, 1952, p. 124).

such as that of Voegelin, put the motive of tension in focus most fundamentally when inquiring about the grounding of the human world.

However, in Schelling's reinterpretation of metaphysical motives, one finds a lack, which comes not necessarily as a mistake, but more so as a result of a larger paradigmatic shift. With Schelling's contemporaries, thought is already on a move towards an immanentist ontology and an anthropological paradigm. This puts in question the capacity of his post-critical thought to relate to its more metaphysical motives in a clear and tangible way (the same issue persistently bothered Kant himself). What seems to be happening here is that the "*spiritual*," seen through the lens of immanent ontology and quasi-anthropology, turns more and more into a metaphoric discourse – one now aimed to express immanent shifts in human consciousness, thus differing from an autonomous pneumatic ontology in the context of transcendent metaphysical hierarchies. Similarly, Schelling thinks of spirituality fundamentally in terms of the way the immaterial appears through the material, but not in light of ideas such as spirit, understood in the classical, pre-modern sense – as a reality with a strictly transcendent axis; this is also what gives way to the tendency for metaphorization. Moreover, since Schelling's thought is invested specifically in the cohesion of its theoretical frame (meaning, in the realization of philosophy as a dominant fundamental science), the lack is most tangible in terms of how such a thought envisions practice: whereas classical metaphysics had clearly defined (and also historically developing – that is, somehow reflexive) forms of spiritual practices – evidenced by the contemplative, theurgic, and magical practices of Platonism (e.g., in Plotinus, Iamblichus and texts such as the Greek magical papyri respectively) – the spiritual in Schelling nears a transposition with the work of consciousness of man in all human practices at all. However, such interpretation doesn't make spirit (in the classical sense) a clearer and, arguably, more accessible reality to us – a clarity that would otherwise be the positive goal and defense of his noetic explication as explication of the metaphysical tradition. Practically, this shift in subject means that there's no actual spiritual practice as such for Schelling, or at least that there's no definite and clear suggestion of what such a practice would entail, if at all possible in the context of the ontologems he puts in use (the new emphasis now being on philosophical/scientific consciousness in method and primarily, though not yet fully explicated, on the anthropological sphere as object). Such an issue doesn't exist solely in a retrograde mind set on thinking metaphysics in a classical fashion; it's an issue of principle for a thought to deal with its own consequences, as posited by its own use of language and theoretical directions. Paradoxically, a pronounced immanentist like Hegel makes a note on the conditions of thinking spirituality by noting how "*Philosophy herself is a worship, is a religion, because she is [...] resignation from the subjective whims and opinions in dealing with the god*" (Hegel, 2010, pp. 40–41).

This resignation at the core of philosophy is, he also points out, the characteristic of spiritual experience, that experience being defined by the turning of the psychic subject into an object in its relation with the spiritual beyond. What's unclear in the case of Schelling is how this transformation is to be enacted as a practice that makes sense of the category of spirit – that is, without metaphorizing the conditions of the transformation.

The paradigmatic shift is demonstrated also through the topic of symbolism. Daniel Whistler observes how for Schelling:

> The universal does not exceed the particular in any way but is entirely identical to it. From this also follows the indifference between finite and infinite, as well as (more obliquely) that of possibility and actuality (623; 196) (Whistler, 2013, p. 147).

This formula addresses the issue of meaning in the production of symbol but also possibly negates such an issue by inherently putting finite and infinite on the same plateau. In this regard, Schelling's thought differs paradigmatically from that of classical metaphysics: the latter clearly situates the infinite above the finite both theoretically (as a noetology) and practically (in its theurgic practices). Using Voegelin's language, we could consider this difference as posited by an immanentist drive at work within the thought of Schelling (which would also explain Voegelin's own ambiguous attitude to Schelling). To give this observation yet another perspective (an epistemic one), Schelling, in opposing classical metaphysics, enacts a scheme in which reason (otherwise understood as *nous* and intimately related to *pneuma* in classical terms) is incapable of operating outside of strictly material reality. This is a claim Schelling's own interest in the esoteric[6] only apparently negates; on the contrary, such interest could also be considered as exacerbating the inner conflicts of his paradigm. For the sake of comparison here, Platonists of antiquity explicitly conceptualized higher (spiritual, immaterial) beings as noetic entities. E.g., even demons were thought of as noetic entities, though of a lower character – such that puts them just a step above the world of sense and, precisely, as reigning in the domain of sense and matter through their downward fixation.[7]

In comparison to other German idealists, Schelling seems to follow a variety of classical metaphysical motives the closest (which defines Voegelin's fundamentally positive attitude towards him). In a post-critical context, he provides those motives with an original interpretation or misinterpretation. The idea of art as necessary for the completion of the broad relation between sense and reason, in the larger context of the grounding of consciousness, fits into such appropriation. This means that Schelling does recognize a *metaxy* at work in the relation between art and understanding by understanding each through a *metaxy* with the ground. On the other hand, he seems to reduce production ultimately to the activity of the Absolute (an issue he continuously grappled with that made him vulnerable to the attacks of other thinkers, as was the case with Hegel's "*History of Philosophy*"[8]). It appears that, while preserving certain language and intuitions of classical metaphysics, Schelling's philosophy focused more fundamentally on conceptualizing immanent relations (under

[6] As certain studies make clear, e.g. O. Pitkänen's "*Schelling, esotericism and the meaning of life*" and S. McGrath's "*Boehme, Hegel, Schelling, and the Hermetic Theology of Evil*."

[7] As laid out in Avramov's study "*The sphere of the demonic according to the period of classical antiquity as a forgotten essence of existential ontology*."

[8] There, Hegel says of the same idea that Schelling's proofs are presented "*in the most formalistic [formell] way, so that in fact they always presuppose what is supposed to be proved*" (Hegel, 1967, p. 267).

the guise of seeking their relation with the Absolute) – such as the relation between art and understanding. I consider this relation as immanent in the sense of it tracing an ontology, or even more so opening up the way for such tracing, in the paradigm of *physis* (the "*physical world*," the "*natural world*," the "*material world*"). Thus, it resonates with contemporary ideas of immanence (such as G. Deleuze's "*pure immanence*") and a more contemporary framing of the question of such relations in terms of the interactions between faculties, social spaces, etc.

5. Conclusion: considering Schelling in light of classical metaphysics

Schelling's push towards immanence isn't a problem by itself. However, I hypothesize that this draw becomes an internal, inherent issue for his thought – particularly in light of his use of metaphysical language, coupled with the eclipsing of classical forms of spiritual theory and practice; for the latter, if my reading is correct, he finds no practical application as such in the context of a living spiritual culture. That is, what is missing from Schelling's view is the characteristically metaphysical/spiritual (otherwise present in the Platonic tradition as a unity of noetology and pneumatology).[9] This approach towards the spiritual and metaphysical from a critical distance, yet seeking engagement through reinterpretation, is characteristically modern. E.g., being a representative of the same German tradition, M. Heidegger sees in philosophy the potential to explicate religious facts (e.g., in "*The Phenomenology of Religious Life*"); however, his position doesn't seem to problematize what it is that a thought defined by a secular drive is capable of contributing to religion, one contribution that would be also particularly religious and that isn't already part of the religious culture as characteristically religious. In other words, whereas in the aforementioned text, Heidegger seems to consider philosophy as introductory to theology, theology already has its own forms of making itself accessible through its own chosen means, e.g., through faith (which Heidegger himself mentions in the same text).

Such doubling of motives between a spiritual and a secular mode of consciousness creates a possibility for further transpositions between the two in this same direction, the doubling being attractive to many as an attempt towards resolving a fundamental difference. However, the theoretical transpositions we discover in late history in relation to this topic seem to be problematic; a better integration of what's transpired in the differing paradigms is work yet to be done. Such work is beyond more simplistic interpretations, both in the conservative realm and by progressivist thinkers like Feuerbach, who reduce all motives of spirituality to a socio-psychological projection (again, the basic issue with this is the destruction of forms of experience and experiencing that are both potential and fundamental for man). It's clear that the problematizing would entail the working out of a fundamental approach that finds the

[9] Such contradictions also seem to be characteristic for modernity. E.g., when it comes to the arts, there's the clear example of Walter Benjamin, who "*criticizes mass culture, but also [...] denies the pretences for autonomy by the modern art and the avantgardes [...] the artwork and artist lose their aura, which leads to the alienation of the public from art and the self-alienation of the artist*" (Kuiumdjieva, 2024, p. 147).

points of relation between the spiritual/religious and the philosophical, scientific, etc., as parts of a singular reality. In this, it makes sense to pay attention to the exemplary Platonic tradition for its own version of the unity of noetic and pneumatic, as well as to turn to thinkers like Voegelin, who offer an authentic modern interpretation of the tradition. Perspectives like the aforementioned present a readymade (for us) positive connection between forms of thinking and practices of immediate grasping (as a form of production – be it interpreted, as it is later by Schelling, in the framework of art, or, as it is with the Platonists, in the framework of spirituality).

When it comes to the topic of art and understanding in particular, Schelling's solution regarding the relation of the two exhibits the traits of a metaphysical framing that, paradoxically, seems removed from the existential and practical situations of its own metaphysical implications. The classical metaphysical culture, in turn, clearly addresses those implications by presenting pathways of spiritual ascent, devotional practice, and forms of vertical communication that rely on a clearly established border between plateaus (e.g., between *physis* and a metaphysical beyond). The presence of the distinctions also allows for a posited pathway of trespassing, of movement between the plateaus (practical forms of transcendence). It's important to note how the contradictions between paradigms are at least somehow resolved within the philosophical "*metaparadigm*" of Platonism and aren't just a clash between philosophy and spirituality. As many researchers of Platonic antiquity know well regarding the issue of grounding, and as clearly seen in Plato himself: "*Rather than a discourse, at its core, Plato's philosophy is revealed as an active dialectic of beauty which seeks a unified vision, extending the feeling of the soul, alienated by the shells and stones that have grown over them*" (Felix, 2016, p. 174).

The relation between art and understanding – in the context of metaphysics, as suggested by the language of the two paradigms and their common ontological emphasis on identity and unity – seems to imply precisely such grounding. I conclude that Schelling's approach to the question of grounding seems to present an issue: first, by its peculiar and subtle aiming towards the logic of the immanent; second, by relativizing the meaning of the Unconditional in his formula for the identity between conditional and Unconditional; third, by not offering a clear practical counterpoint to his thought in view of the classical forms he reinterprets – and in light of the vertical relations they presuppose. As I mentioned, such issues open up Schelling's thought towards a later theory of pure immanence, which is opposed in direction to its historical roots in the late Platonic, characteristically metaphysical framing of the relation between understanding and the immediate symbolic grasp (exemplified by the scheme of theurgy). The contradictions between classical metaphysics and modern reinterpretations of its motives – such that, e.g., horizontalize and/or politicize[10] those motives – become clear through topics like that of the relation between art and understanding. This difference is made especially clear through juxtaposition, e.g. with the classical Platonic tradition. There's a difference between an abstract grounding in an Absolute – as Hegel pointed out in his criticism of Schelling, that is, in an identity between conditional and Unconditional that's simply presupposed – and a living defense of the

[10] As Voegelin's point of focus is in many of his criticisms of modernity.

concrete relationship between immanent and transcendent (which, there are reasons to hypothesize, neither Hegel nor other moderns adequately address as such). This living defense is enacted in a particular form through the theurgical practices of Platonism, as long as they provide an exemplary framing of the symbolic that also makes sense of the metaphysical language of the tradition – thus presenting an actual unity between theory and practice. This doesn't mean the classical approach couldn't or shouldn't be problematized; my sole aim was to uncover the aforementioned contradictions by providing a simple contrast.

References

Bowie, A. (1993). *Schelling and modern European philosophy: an introduction*. London & New York: Routledge.

Croci, G. (2024). The Aesthetic Intelligibility of Artefacts: Schelling's Concept of Art in the System of Transcendental Idealism. *Estetika* 61 (2), 158–175.

Felix, M. (2016). Philosophy and theurgy in the thought of Iamblichus: Symbol and beauty. In Martín-Velasco M. & Blanco, M. (Eds.) *Greek Philosophy and Mystery Cults*. XII, 244 S. (171–185). Newcastle upon Tyne: Cambridge Scholars Publishing.

Fisher, N. (2024). *Schelling's Mystical Platonism: 1792–1802*. New York: Oxford University Press.

Hegel, G. W. F. (1967). *Lectures on the history of philosophy: The Lectures of 1825–1826*, 3. Berkeley, Los Angeles, Oxford: University of California Press.

Hegel, G. W. F. (2010). *Lectures on the philosophy of religion*. Sofia: East-West.

Kassabov, O. (2008). Art and the mirror of philosophy. Reflexion, identity, and system in Schelling 1800-1805. *Altera Academica*, II, 1 (5), 51–66.

Kuiumdjieva, I. (2024). *State planning, culture, and ideology*. Sofia: Faber.

McGuire, S. (2012). Voegelin and Schelling on Freedom and the Beyond. Retrieved February 25, 2026. https://voegelinview.com/freedom-and-the-beyond-pt-1/.

Penev, G. (2013). The tragical myth – towards a symbolic and cosmological interpretation (Lossev, Schelling, Florensky) // *Philosophical alternatives*, 4/2013, XXII, 23–31.

Schelling, F. W. J. (1978). *System of transcendental idealism*. Charlottesville: University press of Virginia.

Schelling, F. W. J. (1989). *The philosophy of art*. Minneapolis: The university of Minnesota.

Voegelin, E. (1952). *The New Science of Politics*. Chicago and London: University of Chicago Press.

Voegelin, E. (2011). *Autobiographical Reflections*. Columbia and London: University of Missouri Press.

Whistler, D. (2013), *Schelling's Theory of Symbolic Language: Forming the System of Identity*. Oxford: Oxford University Press.

Balkan Journal of Philosophy (2026), vol. 18, issue 1, 147 – 174

Book review

Andrei Ionuț Mărășoiu and Mircea Dumitru (eds.). Understanding and Conscious Experience: Philosophical and Scientific Perspectives. New York & London: Routledge, 2024, 258 p. ISBN: 9781032568072

[Reviewed by **Sandra Cătălina Brânzaru***]

1. Why think about *understanding?*

As the title straightforwardly suggests, some of the issues tackled in *Understanding and Conscious Experience: Philosophical and Scientific Perspectives* concern the intertwinement of understanding and conscious experience, but the central topic is *understanding* along its multiple dimensions: most contributors tackle scientific understanding (de Regt, Pritchard, Gurova, David-Rus, Martínez-Ordaz and Macías-Bustos), others explore the phenomenal dimension of understanding, such as grasping (Bourget), apperception and awareness (Elgin), pictorial representation (Stoenescu); some address domain-specific understanding, such as logical proofs (Dumitru), high-energy physics (Martínez-Ordaz and Macías-Bustos), moral understanding (Kelp). The book is structured in three parts, respectively "Part I – What is understanding", "Part II – Understanding and consciousness", and "Part III – Domain specific understanding", and an introduction, where Andrei Mărășoiu thoroughly analyzes the particularities of each contribution, while also performing a state-of-the-art review of understanding.

What does it mean to understand? This is the question that drives the first part, but is pursued in most chapters of the book. There are different ways in which the word is used and it can thus have many different meanings. However, the most prominent in epistemology are atomistic understanding ("wh-" questions/understanding why) and objectual understanding (holistic understanding/interconnections). As Elgin puts it, *objectual understanding is holistic*, it entails grasping relations and interconnections between parts, whereas atomistic understanding (usually) banks on a causal explanation. Some embrace the distinction (Elgin, Martínez-Ordaz and Macías-Bustos), others assume that objectual understanding can be explained using atomistic understanding (Pritchard).

What all authors seem to agree upon is that understanding is *an epistemic good*: we need it to make sense of intellectual accomplishments, especially in science and art (Elgin, 1996). More than knowledge, or at least aside from it, understanding is *a result of success* (Martínez-Ordaz and Macías-Bustos discuss the demands of such success and possibility of actually fulfilling them in cases of genuine understanding). Pritchard thinks understanding is a *strong cognitive achievement* (more epistemically demanding than simply knowing, which only requires weak cognitive achievement).

*University of Bucharest; sandra-catalina.branzaru@fpse.unibuc.ro

Understanding is more than knowing in the sense that there might be something it is like to have the former, it might be brought about by some aspect or activity of consciousness, or as Dumitru argues, bringing cognitive phenomenology to the table, by cognitive and semantic qualia (when one understands *that one knows something* and there is something it is like to be in a cognitive state p). Stoenescu conceives of understanding as a cognitive faculty that comprises *a skill, a process or an accomplishment*, accounting for tacit understanding, and deems it epistemically valuable since it gives us more than "explicit propositional truths", bridging the chasm between cognitive and emotive understanding, and conceptualizing a new way of grasping the differences between understanding in science and in art (emphasizing the role of exemplification and pictorial representation in understanding the phenomenal). Toboșaru hints at how understanding might lead to epistemically transformative experiences, how understanding can be revealing, illuminating it can be not only cognitively transformative, but also *identity transformative* (it can change us as persons, it can change our *holistic doxastic state*). For David Rus, understanding without explanation is valuable because it can account for forms of understanding in scientific practices, offering a richer account of scientific understanding. Gurova defends a minimal inferential account of understanding (acronym: MIAU), which is valuable in addressing the relationship between predictions (including those made by AI) and understanding. De Regt conceives of understanding as a skill, dependent upon history and context, developed in scientific practices, such as constructing theories or models – its value arises from the understander and the context of the understander; it is pragmatic. To understand, one must possess the skill to efficiently make use of models. Kelp defends a systematic knowledge account of understanding. To him, moral understanding is valuable because it is required in order to make sense of what gets to count as good moral action. Bourget advocates for a phenomenal approach to grasping and understanding – the epistemic value of grasping consists in that of phenomenal experience; grasping bridges experience to knowing, allowing for more impactful understanding we can act on, as opposed to simply knowing or believing.

Why think of it as an epistemic good? It must fulfil some epistemic function(s), both objectively and subjectively. And these functions cover more than knowledge. First of all, it fulfils the function of evaluating: explanations (inferences and the relations between inferences), theories, models, cognitive achievements or some other cognitive processes (knowing some statements by heart). It can also help assess the process of inferring (one's own process for example when operating in uncharted, unfamiliar territory, where inaccessible inferences and unrevealed steps give rise to some feeling of insecurity, of disorientation, of incomprehension, when the constituent parts don't seem to cohere – cognitively and/or phenomenally), maybe because we want to avoid confusion or misunderstanding (which would pose a risk to the outcome of what it is to be understood), maybe because the need to understand arises exactly from the feeling of incomprehension – so understanding would be a solution. The question is not what motivates understanding, but what drives us to think about it. Why would we ask, "*Do I (really) understand this?*", "*Did they understand?*", "*Is this a product of understanding?*". Before even articulating these, other questions drive

the quest. Understanding is sought out for predictability, as some sort of confirmation, something that *has our back*: knowledge can be treacherous, inferences too, as well as observations, and explanations. *Grasping* is more secure than *knowing*, and *understanding* may be more secure than *grasping*. To show this, consider tacit and explicit understanding in turn.

Tacit *grasping* is not as secure as tacit *understanding*. A tacit grasp is a mental grip of something yet not articulated, nor conceptualized. It might be revealed in some practice, some skill. But not in conscious awareness. Tacit understanding on the other hand is available in consciousness. Is tacit understanding nonconceptual, like tacit grasping? Then how may it be available in consciousness? It must be conceptual, yet it cannot be articulated. As Janvid puts it, "*to grasp is to access the inference in operation. It is through grasping that the actual access is obtained*" (Janvid, 2018, p. 374). Building on this, explicit grasp is when the inference is revealed in the inferring process, tacit grasp uses that inference without revealing it.

Explicit *grasping* is less secure than explicit *understanding*. Explicitly *grasping* something might be revealed as an ability to talk about it, recognize it, provide explanations (e.g., grasping the rules of chess or whist), but it is only explicit *understanding* that grants you the ability to actually apply the rules and play.

Elgin would agree that tacit understanding leaves one more vulnerable as opposed to explicit understanding:

> Understanding can be tacit. To the extent that it is, the agent relies on it as a basis for inference or action without appreciating that or why she does so. This, I suggested, leaves her vulnerable. Even though she changes her network of commitments in response to new inputs, adjusting it in the face of whatever impediments she encounters, she does so blindly (Elgin, p. 138).

As she further develops it:

> The epistemic situation of someone whose understanding is subliminal is less than ideal. Unsurveyability results in epistemic insecurity, even if the agent is unaware of it. it. When her understanding is tacit, an agent cannot identify the strengths and weaknesses, opportunities and obstacles that figure in her thinking about the topic (Elgin, p. 138).

Maybe an internal state of surveillance directed at the inferring process seeks out understanding. Not thinking, not inferring. Before we *know* we're lost, we surveil our cognition (inferences, thoughts, steps taken in our thinking). We don't surveil our mind unless something *seems off* or if we need to secure something (a theory, a proof, an inference, an explanation, a theorem). Maybe we want to rescue the outcome and identify the error, the gap, especially if the stakes are high. We want to get the pieces hanging together *the right way*. We think that if they do, then we understand. What is the right way? Perhaps a high level of coherence, systematicity, consistency, transparency. A criterion we may all agree upon (for a decade, for a century, for a while, hopefully).

2. Scientific understanding

Understanding has been considered either factive or non-factive, with the debate centring on whether understanding requires holding true propositions or if a false but useful account can still provide understanding. Factive theories, similar to knowledge, hold that understanding necessitates the truth of relevant propositions, while non-factive theories allow for false, but meaningful accounts or idealizations to contribute to understanding, particularly in scientific contexts. A feature of cognitive science is that it uses many idealizations, e.g., "mental representation", "computational theories of cognition" (see Coelho Mollo, 2021[1]; Taylor, 2023[2]; Kirchoff, 2025[3]). Idealizations might be false assumptions, might be felicitous, might elicit true beliefs, might be good instruments for coming up with scientific models. Idealizations *qua* falsehoods may not provide epistemic value to understanding: we need a true or a correct explanation of phenomena in order to get scientific understanding. Elgin would disagree – idealizations *qua* felicitous falsehoods have epistemic value because they might yield understanding in the sense that they train an ability to act, e.g. draw inferences (Elgin, 2022)[4].

According to de Regt, domain-specific understanding might share features with scientific understanding, such as a feeling of familiarity (or grasp of the structure of the phenomena one targets, be that a physical phenomenon or a musical piece). What prevents one from understanding a musical piece or a quantum physical phenomenon could be the lack of familiarity with its language. In his view, being familiar with the *language* (or structure) of what you are trying to understand seems to be a necessary condition for understanding it. The question that lingers is whether it is also a sufficient condition. His contextualist approach allows for understanding at different levels – one who is not familiar with the terminology can still experience, to some degree, understanding of the targeted phenomenon (e.g. a musical piece or a quantum phenomenon) by *observing* its structure as an *intuitive insight* (de Regt, p. 31).

Building on this, where does interdisciplinary understanding fit? In his monograph, de Regt (2017) distinguishes between three levels of science and two methods of arriving at scientific understanding. One can understand a phenomenon, in the sense that one could provide an adequate explanation of said phenomenon, and in order to do that, one must use a vocabulary that refers to accepted items of knowledge. Or one can understand the theory, which means that one would be able to use said theory at a pragmatic level. His three levels of science distinguish between *the macro level* (understanding in science, in general, understanding *why*) – in order to get to this level of understanding, one must be skilful: grasp the interconnectedness of different parts,

[1]Coelho Mollo, D. (2021). Why go for a computation-based approach to cognitive representation. *Synthese*, 199(3), 6875–6895.

[2]Taylor, S. D. (2023). Afactivism about understanding cognition. *European Journal for Philosophy of Science*, 13(3), 43.

[3]Kirchhoff, M. D. (2025). Idealization and mental fictionalism. *Philosophical Psychology*, 1–22.

[4]Elgin, Catherine (2022). Models as Felicitous Falsehoods. *Principia: An International Journal of Epistemology*, 26 (1): 7–23.

how they *hang together*, the *meso* and the *micro* level. Understanding *why* grants understanding *what*. Understanding is then a consequence of applying those skills; understanding at the *meso level* is had by scientific communities (de Regt, 2017:90), for instance producing an explanation that may be *intelligible*[5] to experts (banking on accepted items of knowledge, as opposed to individual idiosyncrasies). De Regt's third level is the *micro* level (individual views of particular scientists). Micro-understanding a theory allows for different communities and different individual experts to accommodate for different theories. Understanding, on this model, is context-sensitive.

On de Regt's framework, domain specific understanding involves using specific items of knowledge, models and practices (these are known by some communities, but not others). So, in order for domain-specific understanding to lead to scientific understanding, the scientific community must satisfy the broad intelligibility criterion, which burdens the provider of the theory to be a skilful understander (an issuer of a theory must be skilled in providing explanations across various phenomena covered by the theory put forward). De Regt's account is contextual, emphasizing the fact that whether the theory is successful or not depends on its applications and predictions, thus allowing for approximate models to provide understanding. Building intelligible theories requires that the issuer should have specific skills. No difference in the nature of understanding *per se* seems to follow from differences in skilled performance, since both theoretical and practical abilities are at play in achieving "success".

3. Transferable skills and integrated explanations

A theory of a particular scientist should be *understood (made intelligible to a community of experts) if said scientist appeals to the* accepted items of knowledge of a scientific community. That community of experts should be able to make sense of how to pragmatically use the theory, and this depends on the ability of the researcher to come up with an explanatory model for it. However, different communities of scientists can deem a theory intelligible or not, based on both context and historical periods. Also, different expert communities might be in disagreement – which raises a serious question for interdisciplinary fields, e.g., cognitive science: "*How can experts communicate their understanding to an audience of non-experts?*" Here, de Regt refers to laypeople as non-experts, but this is also exactly the case of cognitive scientists. How can a neuroscientist communicate their understanding to an audience of computer scientists, anthropologists, psychologists, linguists and philosophers (assuming they form a community of experts in cognitive science)? All the non-neuroscientists may have knowledge of neuroscientific facts, but lack the skills that neuroscientists have. So, the burden is on the neuroscientist to make his understanding intelligible to the mixed community (of experts in different fields), in some sense to transfer "the skills" to the non-expert community, not just the scientific facts. So which skills may those

[5] *CIT1: A scientific theory T (in one or more of its representations) is intelligible for scientists (in context C) if they can recognize qualitatively characteristic consequences of T without performing exact calculations.* (de Regt, 2017, p. 102).

be? Perhaps – the ability to reason with specific concepts within specific theories, given some models?

That seems hard – de Regt's suggestion (in the case of communities of experts communicating with communities of non-experts) is to *employ analogies or metaphors (de Regt, p. 39), which may increase understanding of an unfamiliar (e.g. abstract) domain by mapping it onto a more familiar (e.g., concrete) domain* (de Regt, p. 38), which can, however, lead to misunderstanding and oversimplification. Can experts in cognitive science achieve objectual understanding? Objectual understanding relies on domain-specific knowledge, and at the same time domain-specific knowledge relies on broader scientific principles at play.

The concept of familiarity here plays an important part – if a non-expert immerses in a community of experts, that non-expert could become familiar with the terminology, perhaps also with the practices, but the skills are not guaranteed for actually *doing* research. Interactional expertise is required for interdisciplinary approaches, but it does not grant contributory abilities:

> People who possess interactional expertise but no contributory expertise can fluently communicate and interact with the contributory experts in a field, although they cannot do research themselves (de Regt, p. 36).

Is it sufficient for a cognitive scientist in an interdisciplinary community of experts to be a *spectatorial cognizer*?

> Cognizant understanding of a topic is an explicit awareness of how the elements of a constellation of commitments relate to one another and how that constellation answers to the evidence that supports it (Elgin, p. 139).

It seems that cognizant understanding does not allow for complete scientific understanding. But *why* is that required, and what for? Elgin contrasts the spectatorial stance with the agential stance. The former only allows for what de Regt calls *interactional expertise*; as Elgin puts it, it is a *purely descriptive stance*[6]:

> The spectatorial stance is third-personal. A spectatorial cognizer can apprehend the warp and weft of her tapestry of commitments. She can identify her various commitments, recognize their interconnections and lacunae, locate redundancies and idle wheels, as well as putatively relevant bits of data that do not seem to mesh. She can survey and explore the network. So far, this is purely descriptive. Her epistemic stance is like that of a cartographer, mapping the terrain, identifying and classifying what he finds (Elgin, p. 139).

Perhaps this is not a problem for *weak* interdisciplinary approaches. If the goal of a weak interdisciplinary collaboration is to grasp different aspects of a cognitive mechanism or process, but not to come up with a new model – a novel (interdisciplinary) framework with improved concepts and efficient methodology – then contributory expertise is not required. But that is necessary for strong interdisciplinary science. Is cognitive science strong or weak in this sense? There is no consensus, some (Graham

[6] I leave aside the normative component of Elgin's commitments and focus instead on its bearing on action: it takes skill to live up to one's commitments.

& Bechtel, 1998[7]; Gardner, 1985[8]), as Hohol also points out (Hohol, 2021)[9], believe the cognitive sciences are interdisciplinary in a weak sense, others argue they are strongly interdisciplinary (Núñez et. al., 2021).[10] Do we need a unifying theory about the mind in cognitive science, or should we settle for pluralism (at the level of explanations, models and scientific practices)? Unification, at least *prima facie*, entails an agential stance. Cognitive scientists need to do something about what they find from each other. They need more than *interactional expertise*, they need to take more than the *spectatorial cognizer stance*, the agential stance is necessary since it allows one to *do something about what she finds* (Elgin, p. 142).[11] How, then, could we obtain a holistic approach to cognition – is integration at the level of explanation even possible? What about integration at the level of scientific understanding? What would scientific understanding in cognitive science amount to?

Galli (2024)[12] argues for explanatory integration in cognitive science, but also for the need to integrate local understanding with broader scientific understanding. He contrasts Khalifa's and de Regt's accounts of scientific understanding. De Regt would emphasize the role of scientific models in scientific understanding. The ability to engage effectively with scientific models is key to his contextual account of scientific understanding. Pragmatically, one should be able to use such models in a proficient way, and if one does so, then one understands that specific scientific domain. But it is not obvious how an integration at the level of models would be possible. Khalifa, on the other hand, places a higher emphasis on explanation:

> S has scientific knowledge that *q explains why p* iff the safety of S's belief that *q explains why p* is because of her scientific explanatory evaluation[13] (Khalifa, 2012, p. 12).

[7]Graham, G., & Bechtel, W. (1998). *A companion to cognitive science*. Blackwell Publishing.

[8]Gardner, H. (1985). *The mind's new science: A history of cognitive revolution*. New York: Basic Books.

[9]Hohol, M. (2021). Cognitive science: an interdisciplinary approach to mind and cognition. In B. Brożek, M. Jakubiec,P. Urbańczyk (Eds.), *Perspectives on interdisciplinarity* (33–55), Krakow: Copernicus Center Press.

[10]Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J., & Semenuks, A. (2019). What happened to cognitive science? Nature Human Behaviour, 3(8), 782-791. https://doi.org/10.1038/s41562-019-0626-2

[11]If domain-specific skills aren't transferable, that might then raise the question of whether practitioners of cognitive science could ever have a unified understanding of the human mind.

[12]Galli, G. (2024). Scientific Understanding and the Explanatory Integration in Cognitive Sciences. In: Aldini, A. (eds) *Software Engineering and Formal Methods. SEFM 2023 Collocated Workshops. SEFM 2023. Lecture Notes in Computer Science*, vol 14568. Springer, Cham. https://doi.org/10.1007/978-3-031-66021-4_7

[13]Khalifa, K. (2012). Inaugurating understanding or repackaging explanation? *Philosophy of Science*, 79(1), 15-37.

According to Galli (2024), Khalifa's *Nexus Principle*[14] might be the key for explanatory integration, but he ponders the role of coherence in such an effort: it is not clear how one could achieve *coherence* at the level of explanations across multiple domains in cognitive science. Kvanvig also raises the need for coherence in understanding, not only at the level of explanation. Khalifa does not agree, and believes coherence is not a feature of understanding:

> (...) understanding's connection with coherence is shallow, floating innocuously atop sturdier depths. In other words, coherence is not part of the "core conception of understanding." Similarly, while the "central feature of understanding" is in the neighborhood of coherence, it isn't at home there. On my view, understanding is quasi-coherent: it walks like coherence and talks like coherence, but it does not require a coherentist account (Khalifa, 2016, p. 139).

In this regard, I tend to agree that Khalifa's view is less committal, hence more plausible – other than the fact that it is not clear what one means by *coherence*, it is not obvious what should *cohere* at the level of understanding, what would make scientific understanding different from domain-specific understanding (moral understanding, musical understanding and so on). Perhaps coherence is relative to a model (i.e. what coheres in a model may not cohere in another model).

Pluralism at the level of explanations might afford some fine-grained multi-level account of the tracked phenomenon, which would provide a more nuanced understanding. What does a grand unified theory in cognitive science afford that pluralism does not? Apparently, it's about *coherence*. However, the conceptual boundaries of coherence are not clear enough, so it might be that a grand unified theory might promote *cohesion* instead of coherence. What seems to call for a grand unified theory is not the aspect of interdisciplinarity, but the multitude of frameworks and theories. Khalifa's UBI (Understanding Based Integration) might be relevant here, but integration at the level of models might be necessary as well. Khalifa et al. (2022) only explore explanatory integration for *some* of the models used in cognitive science: mechanistic, computational, topological and dynamical. That would provide at most *partial understanding* or would entail, if aiming for a complete account of understanding,

[14] Khalifa's Nexus Principle refers to "grasping" an *explanatory nexus*, which captures the *correct* relations between explanations and the phenomena these explanations track. Based on this principle, an expert in neurobiology can *understand* some neuroscientific explanation better than an expert in linguistics, for example, if she can grasp the relations between the *correct* neuroscientific explanation and the target phenomenon, thus providing a more detailed understanding, based on which several aspects of the phenomenon can be tracked at different levels of analysis. Solely knowing what a neuroscientific statement is about does not suffice for understanding it; explanations must connect it to the phenomenon. The problem is, sometimes in an interdisciplinary field, the phenomenon being tracked might differ (take, for example, theory of mind or empathy). Explanation might *resemble* scientific understanding, and in order to prevent accidental resemblances, a scientific methodology and processes of evaluating those explanations should exist. Khalifa's SEU (Scientific Explanatory Understanding) serves this purpose; it evaluates explanations. But how can a mixed community of experts evaluate interdisciplinary-specific explanations, how might they achieve the necessary knowledge of the mechanisms or connections that underlie the phenomenon (assuming they are indeed tracking the same phenomenon), and how might they gain the skill to efficiently make use of the specific models (for each domain) to tell if explanations were integrated accurately?

reductions of some models to either mechanistic, computation, topological or dynamical models. It is not obvious how selective integration could provide the fine-grained nuanced understanding advertised by pluralism, nor the fluency or coherence of a grand unified theory.

Barman, Caron, Claassen, and de Regt (2024) make the same move in "Towards a Benchmark for Scientific Understanding in Humans and Machines", where they offer a Scientific Understanding Benchmark (SUB), aiming to measure scientific understanding (for a variety of agents, both human and non-human, e.g., Large Language Models). This approach will be further discussed below in the section titled *Understanding as performance*.

4. The interplay between understanding and cognitive science

Lilia Gurova, Henk de Regt, María del Rosario Martínez-Ordaz and Moisés Macías-Bustos focus on scientific understanding directly. Other authors focus either on domain-specific understanding or understanding in general. Understanding seems to entail some activity unfolding in the *mind* of the understander: *cognitive* activity, *phenomenal* activity, *conceptual* activity. Cognitive science is particularly interested in such activities; sciences of the mind study them, usually in an interdisciplinary effort. It might have been interesting to see whether understanding in interdisciplinary fields, such as cognitive science, qualifies as scientific understanding. Or is it domain-specific understanding? Can we have holistic (objectual) understanding in cognitive science, and if so, what would that entail? Should the explanations be integrated into all models for all domains of cognitive science(s) (even if we don't follow Khalifa in assuming understanding should be explanatory)? Should the explanations and the theories that accommodate them be coherent (what might that mean and how could one measure it)? Should the scientific practices be complementary, should the scientists share domain-specific skills, should they also be able to use domain-specific models pragmatically, effectively across all other domains? Would understanding what cognition is without having the required skills (or knowledge, or model) be sufficient to count as understanding "cognition"?

There are two main ways in which understanding might be relevant in relation to cognitive science: what gets to count as scientific understanding in cognitive science and what understanding is from a cognitive science perspective.

From the perspective of cognitive science *understanding* can be conceived of as:

i a process

ii a product (of a process or of multiple processes) (first person approach)

iii a performance (third person approach)

Understanding can refer to a process or to the outcome of a process. If the outcome of a process is an ability, then we will treat that separately, as evinced in performances (third person approach). If understanding is a process, then first of all,

we must differentiate it from other processes. Is it different from cognition? If so, in which way(s)? It might seem that understanding is dependent on cognition: if we think of it as a "cognitive achievement", it entails doing something with some content by thinking about it. If it is the result of cognition, then it cannot be a process: it is some sort of *outcome* – if that is the case, then understanding is a product of some sort of cognition. If it is a process, but we maintain the assumption that it is related to cognition, then it must, at least in some sense, depend on it or be related to it. How do they relate? Is cognition necessary for understanding? If this is the case, then we cannot have understanding without cognition. Both concepts are somewhat mysterious, very much like creativity. We ascribe cognition, we ascribe understanding (to ourselves, to others). Cognition is also ambiguous; there is no clear or unanimously accepted definition of what cognition is. The attempt here would isolate what the central features of both cognition and understanding may be.

The classic approach in cognitive science is that cognition is based on operating upon (via procedures, usually thought of as computational) some representations (or representational structures). There is sufficient disagreement in the literature on what representations are and what computation is to be sceptical about a model of cognition. However, this broad description of what cognition may be encompasses most theories wielded in cognitive science (symbol system hypothesis, connectionist models, artificial neural networks). This view has faced many challenges: the idea that minds function based on representation and computation might be mistaken, either because it neglects the role of the body, or the physical environment in which the body is situated, the way in which some *mental* processes extend into the world, the role that consciousness plays (the phenomenal aspect of consciousness), the role emotions play, the role that society plays, the interdisciplinary aspect of the field itself and the inability to develop a core, unified theory that would cover all its disciplines.

In light of all these challenges, a novel approach emerged: the E-cognition approach to cognition (usually referred to as the 4E approach). The central hypothesis of E-cognition is that mind is a dynamical system (Haken, Kelso, and Bunz, 1985[15]; Van Gelder, 1995[16]; Smith and Thelen, 1994[17]) or an *enaction* (enactive cognition) (Hutto and Myin, 2013[18]; Di Paolo, 2018[19]; Menary, 2006[20]; Noë, 2014[21]).

[15]Haken, H., Kelso, J. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological cybernetics*, 51(5), 347–356.

[16]Van Gelder, T. (1995). What might cognition be, if not computation? *The journal of Philosophy*, 92(7), 345–381.

[17]Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. MIT Press.

[18]Hutto, D. D., Myin, E. (2014). Neural representations not needed – no more pleas, please. *Phenomenology and the Cognitive Sciences* 13, 241–256. https://doi.org/10.1007/s11097-013-9331-1

[19]Di Paolo, E. A. (2018). The enactive conception of life. *The Oxford handbook of cognition: Embodied, embedded, enactive and extended*, 71–94.

[20]Menary, R. (2006). Attacking the bounds of cognition. *Philosophical psychology*, 19(3), 329–344.

[21]Noë, A. (2014). Concept pluralism, direct perception, and the fragility of presence. In *Open Mind*. Open MIND. Frankfurt am Main: MIND Group.

"Cognition" does not usually refer to a single process, and it is not conflated with reasoning: there are basic kinds of cognitive process, complex cognitive processes (the threshold between what gets to count as basic versus complex is not clearly defined, however, we will work with the differentiation in a broad sense). Some examples of cognitive processes, usually deemed as basic, are perception (interpreting sensory information), attention (assigning mental resources to a specific thing, be that an activity, a thought, or a specific perceived input), and memory (encoding, storing and retrieving some content). There are many cognitive processes deemed *complex*: reasoning, language use, decision-making, metacognition, learning, critical thinking. The distinction seems to be based on the obviousness of the use of concepts: basic cognition may allow for nonconceptual aspects to play some role, but complex cognition heavily relies on the use of concepts.

5. Understanding as a process

Although all of them may be involved in understanding, we want to isolate the cognitive processes on which understanding seems to heavily rely: perception seems to be a good candidate (Stoenescu makes a good case for the role of perception in understanding in *Ways of understanding the phenomenal*, via pictorial representation and exemplification), memory as well (Bourget's account of grasping clarifies the role of memory, by distinguishing between merely *thinking* about something and truly *grasping* it; for Bourget, memory plays some role in cases of grasping, e.g., when Mary fetches experiences *from memory as needed to ground her thinking about colors* (Bourget, p. 115)). Attention seems necessary for evaluating explanations, but it plays a great part *in absentia*; think of "aha" moments – what makes them special is the fact that these appear when you are not *paying attention* to the subject matter, when you are not attending to the problem, but rather to something else. Focus seems required when thinking about the problem, and it is also important in the conceptual articulation necessary for making sense of the *aha moment* – its absence (or *not attending* to the problem at hand) seems necessary in the *transition* or *incubation* period before the "aha" moment, and in that moment, attention resurfaces in order to capture the *feeling* of the breakthrough and access the content of the breakthrough. Added to these three basic cognitive processes we might need "complex" ones as well: metacognition (for evaluation) and *inferring*.

Inferences probably have the best seat at the table of understanding. These result from inferring. Inferring is something that agents (organisms or machines) *do* when they draw conclusions from premises, often going *beyond* a given input (data, information). Inferring is considered key in *extending knowledge*: so how does that happen?

It seems that first one needs to isolate some "knowledge", some "representations"[22] from the premises (it is not obvious how) and then one tracks those representations while doing some operations[23] (usually operations that are described by appealing to certain types of logic; it is not obvious if what is described matches the actual operations) until those representations re-organize, based on the operations, in a different form, usually a conclusion (there are footsteps in the sand, they have the shape of shoes, so someone must have been here) or in confirmation of one of the hypotheses as "being the case" (p or q, not q, hence p)). Based on how they are carried out, inferences can be conscious or unconscious (Reverberi et al., 2012).[24]

What is the relation between inferring and understanding? Often, inferring brings about more or a novel kind of knowledge, or it settles some competing options for what actually is the case, but most contributors agree that understanding is not reducible to knowledge (Elgin, Pritchard, David-Rus, Bourget, Stoenescu, Dumitru). De Regt argues that scientific understanding would primarily be a kind of knowledge, and Khalifa would reduce understanding to knowing. As Paul Boghossian points out in his 2014 paper, "What is inference?", the topic is understudied by philosophers, but also by cognitive scientists. Assuming understanding is different from knowledge, it seems important to explore the role inferences and the process of inferring play in understanding. Gurova carefully outlines a minimal account of understanding where inferences are central:

> MIAU (core statement): The person S understands P (where P could be a theory, a phenomenon, or a linguistic expression) if and only if S can produce correct inferences from the knowledge which S associates with P (Gurova, p. 69).

And here are the corollaries:

> Corollary 1: The more (non-trivial) inferences S can draw based on her knowledge of P, the deeper her understanding of P (Gurova, p. 69).

It seems the process of inferring is tied to understanding based on the outcome of the process, which consists in an inference that can have the form of a *realization* (statement, conclusion, opinion), which is then evaluated (by a different cognitive process, perhaps metacognition?) based on some criteria (perhaps simplicity, consistency,

[22] I use "representation" in a broad sense. In a 4E approach, inferences are not viewed as some internal symbolic computations. These are derived from basic interactions between the agent, its body, the environment, and other agents in the environment. On an enactive account of cognition, inferences are active, as opposed to static and representational. In the traditional approach, inference is based on "collecting" passively some input and then processing it, in the sense of performing some computation based on some rules, whereas on the enactive approach, *active inference* not only collects some "input" or "data", but rather it *does* something – as Friston put it, "*the very fact that you are reading these lines means that you are engaging in Active Inference – namely, actively sampling the world – in a particular way – because you believe you will learn something*" (Parr, Pezzulo and Friston, 2022, vii). In this sense, active inference is a means of reducing uncertainty, and plays a functional role in hypothesis testing or making predictions.

[23] E.g. rules of inference: conjunction, disjunction, Modus Ponens, Modus Tollens, hypothetical syllogism.

[24] Reverberi, C., Pischedda, D., Burigo, M., & Cherubini, P. (2012). Deduction without awareness. *Acta psychologica*, 139(1), 244–253.

coherence). From a cognitive perspective, what is the difference between premises, inference, conclusion and explanation? Do they differ based on their epistemic functional role; do they differ based on the cognitive processes that produce them?

> Corollary 2: An explanation of P increases S's understanding of P if and only if it allows S to make additional inferences (Gurova, p. 69).

What is the difference between explanation and inference? Is the explanation not another inference that, based on some requirements, gets to count as *explanation*? Cognitively, how are they different? Is it the content (coming into a new form)? Is it just the function (it connects a block of inferences)? Is it attitude? ("I can now endorse this", "I can now accept this"). Functionally, it seems that premises are tied to input: the inference processes the input (based on some rules), the output is the conclusion, and the explanation bridges (logically or causally) the premises to the conclusion (the input to the output). The explanation is then another inference or a result of metacognition (one thinks about why they thought that was the conclusion and how they reached that conclusion). But what does metacognition add, epistemically? Perhaps at the level of explanation, the process of inferring becomes transparent, one makes inferences about the initial process of inferring. Would those inferences about the already made inferences be additional inferences or explanations? What are additional inferences? How does one differentiate them in order to count them? Are all of them transparent? They should be if one is working on a proof. But what if the logician relies on some unstated assumptions that she is not familiar with (think biases, think inherited tacit assumptions that pertain to a particular logical framework rather than another)? Familiarity here could be felt, especially when working on proofs, but as Dumitru points out, this does not guarantee we got it right:

> Granted, having the feeling that the proof is correct or more generally that we have got things right does not guarantee the logical validity and soundness of the proof or of other epistemic enterprises. Nevertheless, the feeling that we got things right helps us with a strong and clear understanding of the epistemic maneuver (Dumitru, p. 181).

But perhaps the feeling of familiarity does not have much to do with actually getting it right, but it helps reveal some aspect of the inferring process. If understanding is a process, what of misunderstanding? The difference between the two could be at the level of content: perhaps the content of an inference, in the case of misunderstanding, is not the "right" one – it is an unfamiliar one, it is different in a sufficient way to not cohere (or do not connect in a way that would allow all the inferences to stay together as a whole) with the rest of the content produced throughout the process of inferring.

Incomprehension might feel like *disorientation* (e.g., getting lost in a city), where there is a disconnect between one's internal map versus the external world. The feeling of incomprehension or lack of familiarity could signal some sort of a disconnect between the perceived outcome of one's inferring process and some part of the process, including the rules and steps based on which the proof unfolded, or a disconnect from one or multiple inferences to which one has no access or which do

not seem necessary. It could be that a mechanism different from the inferring process (perhaps memory, perhaps intuition or some unconscious process, like pattern recognition) handles the content of what seems to the reasoner to be inferences.

When some content (path, building) appears familiar, one might feel like "*I know where I am*", "*I recognize the place*". This does not entail that the person now knows exactly what the path is, but they are no longer lost, and this might be sufficient to stop the feeling of incomprehension. But lack of incomprehension does not necessarily entail comprehension. However, the high contrast between no longer making sense of where I am – complete disorientation versus finding something that I have the feeling is familiar – may make it look as if "having an idea of where I am now" means "understanding where I am". It may seem that this is what grasping actually captures: that phenomenal aspect of familiarity triggered by some cognitive content that has not yet been *recognized* (has not yet matched any representation). What would the relation between grasping, inferring and understanding be if the three are different cognitive processes? We could perhaps delineate grasping from inferring – grasping might be procedurally different, triggered when inferring is not transparent enough to the reasoner.

Strevens would argue that grasp is *some relation between the mind and the world* (Strevens, 2024, p. 745)[25], and that it *manifests itself in a kind of inferential ability, and thus, that a lack of grasp can be diagnosed from certain inferential limitations* (Strevens, 2025, p. 4). So – not a process, but an ability, an inferential ability. In relation to understanding, he argues that "*various forms of understanding differ in what is grasped, but not in the nature of that grasp*" (Strevens, 2024, p. 742). His ability-based approach makes understanding dependent on grasping, but it looks like understanding is *a matter of the mind's encountering, framing, connecting to, in the right sort of way, some aspect of the world. Understanding is a place where the mind meets the world* (Strevens, 2024, p. 742). To understand something one must grasp an explanation for that thing.

If we think that understanding is a process, then the grasp of an explanation might be the outcome. If we think that understanding is an outcome, "understanding" may mean having grasped an explanation. If an explanation is the result of inferring, then a grasped explanation could be the result of a skilled process of inferring (inferring with grasp-ability).

On the other hand, Bourget argues for a phenomenal approach: *grasping some mental content is a matter of having it in phenomenal consciousness (we grasp to the extent that our thoughts are grounded in experience, whether occurrent or non-occurrent). This is not a semantic point about how we use the term "grasp" and cognate expressions. This is an empirical claim about how thinking works, and one that has important downstream implications because it explains why grasping and experience both seem to matter in cognition* (Bourget, p. 126). If that is the case, and grasping is necessary for understanding, then the process of inferring is

[25]Strevens, M. (2024). Grasp and scientific understanding: a recognition account. *Philosophical Studies*, 181(4), 741–762.

insufficient for understanding, phenomenal consciousness is required and involved in our cognitive processes.

> Corollary 3: Explanations (i.e., structures such as "P, because Q") are not the only means for achieving understanding, other structures that enable inference (predictions "Given P, we should expect Q", categorizations "P is an instance of Q") can do the same job (Gurova, p. 69).

What, then, is the difference between a prediction and an inference, other than, grammatically, the fact that they answer different questions? A prediction answers 'what', an inference answers 'why'. Time orientation? Prediction, at first glance, seems to be about inferring the future. If one is trying to make sense about what happened in the past, would the *prediction* turn into an *inference*? (E.g. in historical or scientific contexts: trying to retrodict how the universe formed, trying to account for what a historical figure did in the past, so as to reconstruct it).

If we are to continue to explore relations between inferring and understanding, it is unobvious how the two relate. On a minimal inferential account of understanding, it seems that the distinction between the two (as processes) becomes irrelevant. If we take understanding to be an inference about the inferring process, that qualifies the target process as "understanding", "misunderstanding", then understanding is not a process, but a quality or a property of the inferential process under scrutiny.

So far, it may seem that understanding is multiply realizable, can either be realized by inferring or by grasping or by some other cognitive activity, but none seems sufficient in capturing understanding entirely, independently: some cognitive activities seem necessary, but they don't work in isolation. Perhaps so as to identify which inference, which skill and which cognitive process, one must look into what the object of understanding is. These might be contingent on the object of understanding. Inference and skill need not vary from individual to individual if we accept there are degrees of understanding. This is why I next consider understanding as an outcome – as that through which the target object or phenomenon is understood.

6. Understanding as an outcome

If understanding is an outcome, we should explore the kind of outcome it is and the cognitive activity (or activities), the outcome of which it is.

- Understanding as some sort of content [articulated, not tacit]: propositional knowledge (or explanatory understanding): *I understand x* (objectual) (where x is an object or a subject matter – philosophy, geography, my husband, physics) or *I understand why x* (atomistic). So, the outcome is some array of conceptions that are articulated in a *coherent* manner, in the form of some explanations or some sets of explanations that hang *together*;
- Understanding as tacit: tacit grasp, tacit knowledge, tacit ability (know-how, procedural);

- Understanding as a quality of an explanation or a mode of presentation of an explanation (one shows full understanding via a way of presenting some explanation);
- Understanding as mental state (I know I understand; I understand what I know, I understand that I don't know, I know I don't understand);
- Understanding as an acceptance – rational credence at a threshold that is minimal or optimal for outright belief[26];
- Understanding as an experiential state (I am experiencing understanding that/why p);
- Understanding as a quality of an experiential state (quality of the experience – I finally *get it, I understand*);
- Understanding as apperception;
- Understanding as updated knowledge, updated sets of commitments (could collapse into an inference, or a set of inferences, which could further be articulated into an explanation, sets of explanations or compounded with other articulated inferences in different forms);
- Understanding as an ability (first person, some cognitive capacity)
  - i an ability to follow an explanation why p;
  - ii an ability to follow one's own inferences and track them;
  - iii an ability to manipulate one's own representations or models of some phenomenon;
  - iv an ability to manipulate explanations (in writing, to oneself; in introspection);
  - v an ability to explain p in one's own words to oneself;
  - vi an ability to draw from q the conclusion that p.

Let's assume it is an achievement, take for example Pritchard's account, where understanding is a strong cognitive achievement: j) [understanding as strong cognitive] *achievement is success that is primarily attributable to one's manifestation of relevant ability and which in addition involves either (i) a relatively high level of ability or (ii) the overcoming of a significant obstacle to one's success* (Pritchard, p. 49). However, understanding in this case is not reducible to knowledge, since jj) *understanding is essentially an active form of cognition, in contrast to knowledge, which can be passive* (p. 53). If understanding is an active form of cognition, then it would be a process. Pritchard also states that jjj) understanding is specifically concerned with the subject exhibiting a *strong cognitive achievement*. If it's about the subject exhibiting some sort of ability, skill, then understanding would be a specific kind of performance that presupposes some skill (act by showing some markers that reveal the relevant ability) or some predisposition.

[26] I thank Andrei Mărășoiu for unpacking "acceptance" this way.

If understanding is reducible to knowledge (which is not a view defended by any of the contributors to Mărășoiu and Dumitru (2024)), it must be some kind of knowledge; if it's knowledge of an explanation, then it should be the outcome of some processes involved with the production of an explanation. Being able to draw inferences seems necessary for producing an explanation. The explanation would reveal the level of understanding. How? There must be some sort of "coherence" or "consistency" which should be visible at the explanation level, not at the propositional level.

If understanding is a product, who must pick it up? Both the one who produces it (the reasoner) and the one who picks up the explanation externally (a community of experts, a student or a non-expert). There might be some match between how understanding was achieved mentally by the internal reasoner and how its output will be unpacked, processed, repacked to achieve understanding by the external reasoner, but that symmetry must be realized at the level of processing. Different skills and methods can be used based on the competence of the external reasoner: the external reasoner must acquire skills for decryption, the internal reasoner must acquire skills for expression, so as to make the decryption level appropriate for the audience. Grasp plays an important part here, the less experienced the audience is, the more refined their level of grasp needs to be (more things need to be grasped: explanations, only broken down into smaller graspable inferences), the more experienced an audience is the less refined their level of grasp needs to be (fewer things need to be grasped: blocks of explanations are recognized, and thus they need not be broken down into smaller pieces that need to be grasped).

7. Understanding as performance

Conceiving of understanding as performance uses the Turing move: not asking what understanding is, but evaluating it from the second- and third-person perspective. It is a behavioural conception, not a novel approach, commonly used in education: we evaluate the understanding of students, our own, we evaluate the understanding of researchers. So we have different kinds of tests, different kinds of "Turing games". These usually involve asking questions, providing explanations, solving problems, giving examples, generalizing some principles. We want to evaluate *understanding*, not knowledge only, so we should make sure our tests are designed in such ways to tell apart simple memorization, Gettier-style cases or misunderstanding from *genuine* understanding. If it looks like genuine understanding, talks like genuine understanding, is it understanding? This question is not useful here.

If we conceive of understanding as performance, then we must have some markers for it, for instance:

- One should provide some explanations, especially in scientific understanding;
- One should be able to transfer knowledge;
- One should have (know) the right models, but also the ability to use them;

- One should know their concepts, but also have the ability to analyze them, use them, explain them;
- One should be able to "draw inferences, raise questions, frame potentially fruitful inquiries, and so forth" (Elgin 2017, p. 3);
- One should manifest familiarity with what is understood (think musical understanding in de Regt's example, where understanding can be "the result of familiarization, but not only a question of becoming familiar) (de Regt, p.34);
- One should manifest "interactional expertise" (de Regt, p. 36) (to understand something in order to manifest an ability to transfer knowledge about that thing), but to contribute one needs also to *construct*, *use* models, to use skills specific to the scientific domain and to acquire those in scientific practice;
- One should be able to grasp in order to understand, for Bourget, *grasping some mental content is a matter of having it in phenomenal consciousness* (Bourget, p.99) *a;
- One should be able to have *conscious awareness of the constitution and contours of one's network of epistemic commitments. It also requires conscious awareness of one's own powers, limitations, affordances and resources – that is, recognizing oneself as an epistemic agent engaged with a particular topic* (Elgin, p.147) *b;
- One should have modes of symbolizing, in order to understand the phenomenal, not only for describing or depicting, but also the ability to sample, where samples which symbolize properties are possessed by both object and *sample* (Stoenescu, p. 166). *The understanding of a pictorial representation requires a background knowledge of pictorial conventions, referential connotations and contextual elements* (Stoenescu, p. 165) *c;
- For some domain-specific knowledge (e.g., logical proofs) one might experience something it is like to be in that cognitive state (cognitive qualia), especially since "the phenomenon of *seeing* something as an aspect of something *else* is instrumental in bringing about a moment or a process of illumination (the so-called *aha*-experience), which is associated with a reasoning process and with the person who is doing the reasoning" (Dumitru, p. 191). That is not strictly necessary for all kinds of understanding, as "aha" moments are not necessary for understanding. In some domain-specific knowledge, however, the exploration of uncharted territory, some threshold between thinking a proof is "nonsensical" versus "it's very novel, unfamiliar, but it makes sense", may be guided by some cognitive/semantic/logic/geometric qualia;
- For genuine understanding, one might need to meet some subjective criteria (transparency of the understanding process, grasp) and some objective criteria (structural character, order, coherence, empirical adequacy) (María del Rosario Martínez-Ordaz and Moisés Macías-Bustos, p. 196).

If we conceptualize scientific understanding as a cluster of some abilities, then we should be able to observe the expression and manifestation of those abilities and,

perhaps, be able to measure them. From those listed above, we must exclude a*, b*, and maybe also c*. These are not ruled out because of their explanatory power, but because of their not being *benchmark-friendly*.

This is what Barman, Caron, Claassen, and De Regt aim to do with their SUB account (Scientific Understanding Benchmark). They follow De Regt's approach and select the following abilities for benchmarking:

> we start from the idea that scientific understanding involves the ability to provide explanations within a theoretical framework that is intelligible to the agent, which includes the abilities to derive qualitative results, answer questions, solve problems properly, and extend knowledge to other domains or levels of abstraction. We argue that various degrees of scientific understanding can be measured by measuring different levels of ability, such as having access to relevant information, the ability to provide explanations, and the ability to establish counterfactual inferences. These different levels can be quantitatively evaluated using what-, why-, and w-questions (Barman et. al., 2024, p. 6).

Behaviorally, we should observe following:

(1) the ability to provide explanations within a theoretical framework that is intelligible to the agent;

(2) (Which includes) the abilities to derive qualitative results, answer questions, solve problems properly, and extend knowledge to other domains or levels of abstraction;

And measurement should account for (the performance of understanding) at various degrees:

(3) by measuring different levels of ability, such as having access to relevant information, the ability to provide explanations, and the ability to establish counterfactual inferences. These different levels can be quantitatively evaluated using what-, why-, and w-questions.

Here are some potential problems that could stem from benchmarking *understanding* and measuring it:

- It is not obvious that understanding is an outcome, or a process or an ability, maybe the outcome and the internal processes and the abilities, competences and the experiential aspect play a role in what provides *understanding*. We could be making a category mistake.
- Distinguishing between the appearance of understanding versus partial understanding – a performance could be evaluated as S *understands* or *lacks understanding*. Partial understanding could be missed.
- Distinguishing between surface-level understanding versus deep-level of understanding – a performance could be evaluated as deep-level understanding or a deep-level understanding could be evaluated as surface-level understanding, because of different failures (articulating an explanation, the ability of the

evaluator to grasp an unarticulated explanation, misunderstanding of a query, ambiguous prompt, so on).

- Assuming that the evaluator has a neutral predisposition and no bias towards the AI.
- Assuming that a human expert has the ability to actually evaluate an AI. Maybe the way in which an AI produces explanations depends on the skill of the expert to actually prompt the correct answers. Perhaps the AI has a different disposition – especially if it is a large language model: instead of aiming for the right explanation the LLM is meant to keep a user engaged.
- A tacit assumption here is that a large language model actually behaves as an interlocutor – is it really an interlocutor? There is no consensus on whether sensory grounding, intentionality, consciousness are all required for *knowing* something to be the case, for understanding something to be the case. We need not dismiss this view, but why take it for granted that they actually are interlocutors?
- Which agent is actually evaluated? Whose scientific understanding is evaluated? For example, Chalmers distinguishes between models, instances and conversations with chat bots, where interlocutors may be *entities tied to models, instances, or conversations. On one hypothesis, there is one interlocutor per model. On another, there is one interlocutor per instance. On another, there is one interlocutor per conversation* (Chalmers, 2025, 8).
- According to Elgin (p.147), understanding is agential: how can a large language model be aware of their epistemic commitments? What would it mean for an AI to be "*recognizing oneself as an epistemic agent, being cognizant of the constitution of one's network of commitments*"?

If we take consciousness as neither necessary nor sufficient, this will however not rule out experience as necessary or sufficient. If we take scientific understanding to be *a skill-based capability that relies on an agent's ability to perform specific actions* then perhaps some clarifications with regards to agency and action are required, provided that an organism's external world is different from that of an AI system. For example, do such systems have linguistic competence? What are their conceptual representations grounded in? Do they inherit grounded meaning (in the external world) from us? Does their indirect relation to the external world bear on their inferential norms? If we aim for scientific understanding of some physical phenomena, then the target is typically environmental, whereas bots deal in code. Perhaps both biological and non-biological intelligent systems, on a functional account, know things, infer and understand, but they do so differently.

This might matter only if we want to integrate their knowledge or understanding with ours. Froese identifies a possible dilemma, which applies to enactivism, but can be further extended: *either frontier LLMs are capable of sense-making despite lacking biological embodiment, or the kind of linguistic competence they exhibit does not necessarily require sense-making* (Froese, 2025, 1). Froese sides with the

latter: LLMs exhibit a novel kind of sense-making. If we extend that to understanding, then either LLMs are capable of scientific understanding despite lacking biological embodiment, or the kind of linguistic competence they exhibit does not necessarily require understanding. Here multiple approaches can bear on the choice:

(1) On a functional account of understanding, biology does not play much role, understanding is multiply realizable, what matters is having the right computations, processes and internal structures that are interrelated (wired) in the right way.

(2) On a behavioural account of understanding, biology does not play any role, but nor does the internal architecture. If it acts like it understands something (has some abilities to pass our tests), then it does. What kind of machine can play the imitation game? The Turing manoeuvre can be applied to the SUB benchmarking attempt – what kind of agents are subject to testing: why not nonhuman animals, why not a social robot, why not a smart vacuum cleaner, why not a self-driving car, why not a bee, why not a crow, why not a baby?

(3) On a phenomenal account of understanding, biology may play some role (sensory qualia), but not necessarily if we accept the idea of cognitive qualia.

(4) On an experiential account, experiencing is necessary. So far experiencing has been thought of as a feature of living organisms. We could accommodate other kinds of entities, but on which grounds? Interpretability may not help here.

(5) On an embodied account, biology does play some role, that of symbol grounding. If we could work out, for an AI, some connection to the external world (an environment co-inhabited with non-living organisms as well).

(6) Another problem is transference of knowledge, especially in the case of new discoveries: assuming that an AI could actually produce a scientific discovery, how will that be transferred? Via an explanation? Before arriving at that point, perhaps the scientific novelty cannot yet be articulated in some common vocabulary (one both the AI and the human interlocutor share). Transference of knowledge may be a different skill. Interactional expertise does not grant contributory expertise. Consider an AI on the verge of producing scientific novelty: the inferential process is unfolding. Transparency is met, and so the agent tracks the inferences. The most important ones are basic for the AI. However, these may not be basic for the human at all. The AI does not articulate all inferences; it only articulates what seems necessary. Now, the evaluator will ask further questions, and the more inferences are revealed the less intelligible it seems to be to the evaluator. How is a hallucination differentiated from a completely novel discovery? Alston's distinction (Alston, 1993) might help, between *access to the ground and access to the adequacy of the ground*. If all goes well, whatever justificatory structure (adequacy to the ground) the AI has, that must reach the mind of the evaluator. In order to do that, reorganizing, rephrasing that structure and re-adapting it to fit the background of the evaluator is necessary. If the jump to the evaluator's background is too big, and too many gaps need to be filled, challenging most dispositions the evaluator

> has for believing something is the case (and the new framework showing that the evaluator had *prima facie* reasons for believing that to be the case, but not *ultima facie*), that would probably lead the evaluator to dismiss the entire ordeal as a hallucination produced by the LLM or AI. Or it can go the other way around too. The problem with transferring knowledge in this case is that the transfer will either be incomplete and knowledge will be corrupted, thus rendered useless, or the recipient party might *prima facie* find it unjustified (because the theories and models operating in their mind will deem it so), but never actually get to finding it *ultima facie* justified in their lifetime.

Another important aspect that Elgin points out is *being cognizant of the constitution of one's network of commitments*. Consciousness aside, transparency is not granted. Other than the fact that it is not obvious "who" the interlocutor is in the case of LLMs, its processes are only partially (and at surface-level) transparent: the network of commitments is limited to the conversation, some rules (safety over completeness – which blocks access to weights, activations, transformers), and the token-by-token generation of output (text in conversation). It cannot inspect its weights, it cannot override its architecture, it cannot self-debug (why this token and not another token). Usually, we don't have access to our processes either. This is what makes *understanding* spellbinding – grasp is some sort of magic power which allows us to catch something in the moment (while the process is unfolding).

Approaching understanding as a performance may appear useful, but it depends on the scope. As Bourget points out, just because some other type of entity performs *understanding*, this does not mean we should not infer that some processes, or consciousness, does not play any role in how we understand:

> it would be a mistake to infer that we don't use our phenomenal states for reasoning because AI models don't use such states. The fact that we can build machines with input-output signatures similar to those of our brains does not mean that what is in our brains is doing no work in generating our outputs. We might as well conclude that calcium ions are doing no work in neurons since AI models don't have those either (Bourget, p. 126).

On the other hand, we should not ignore the potential benefits of exploring the features and processes underlying the performance of AI systems. In order to understand how their performance impacts epistemology of understanding (and what we can learn which might prove useful in delineating the cognitive processes at play in how we understand something), we should first try to outline the differences, not reduce them to performance. I believe that Froese makes an important remark when analysing his "AI dilemma":

> to properly understand how distinct domains can relate to each other in a reciprocally efficacious manner, we must first learn to properly separate them (Froese, 2025, p. 11).

8. The role of consciousness and experience in understanding: coda

Epiphenomenalists and illusionists would blatantly dismiss any role consciousness may play in understanding. Understanding is generally construed as something that happens *from within*, within the *mind*, so it may be plausible to think that it is something specific to organisms that have minds. Usually we use it in relation to humans (John understands how the car engine works, Mary understands what the hard problem of consciousness is, Andrea understands how to play whist, my neighbour understands Rembrandts' "*The Night Watch*", I have never understood the "*Hammerklavier*", Maddy understands Icelandic, I understand the legend "*Jólakötturinn*", Lydia understands John's anger, Gowers understands the mathematical proof, the child understands Snow White's concern), or some non-human agents (e.g., my dog understands that I'm tired so that's why he's not barking, or my dog understands why we're not playing – he's punished because he chewed on the couch) and (recently) with large language models (BERT understood the assignment, or ChatGPT understands me better than my friends).

Some may think understanding is contingent upon consciousness – it doesn't necessarily require it to occur – although it's important to factor in the multifaceted concept of consciousness: if we use the word to mean conscious organisms, as Elgin points out, it is likely "trivially true" for it to be a prerequisite for knowing or understanding anything. It's not obvious, though, that this would prevent us from using the word in relation to fictional characters (e.g., Santa Claus understands that I've been better this year), institutions (e.g., the Government) understands our requests, or large language models (e.g., Qwen understands me better than my friends).

If by consciousness we refer to its phenomenal aspect, then the relationship is nuanced: understanding might be something that one is conscious of, there might be *something it is like to* understand (see Dumitru or Bourget or Stoenescu). One might think that consciousness plays a central role in understanding – e.g. consciousness is necessary for grasping, like Bourget: *grasping some mental content is a matter of having it in phenomenal consciousness* (Bourget, p. 99). Yet, consciousness may not actually be instrumental in understanding – according to Elgin, *whether or not there is a felt quality to conscious awareness makes no epistemic difference* (Elgin, p. 18). Perhaps the felt quality makes no epistemic difference at some explicit level, because other processes can do the work. But what if those processes are not available?

> When we get the idea of a proof, when it clicks and we can now follow and produce similar proofs with ease, gaining perhaps also the ability to explain it to a fellow student, it would seem (to me, at least) as though cognitive qualia are at play (Dumitru, p. 181).

When working on a proof, perhaps the reasoners understand something they cannot yet articulate, or know something but don't understand how they know it. Or, perhaps they understand something without understanding that they understand. Zagzebski would simply dismiss the latter:

> It may be possible to know without knowing that one knows, but it is impossible to understand without understanding that one understands (Zagzebski, 2001, p. 246; also in Janvid, 2018, p. 374).

The two, for some, are actually not the same epistemic state. The relation that p stands to q could be taken at face value, without accounting for the inferential relationship. Grasp is important here, as Janvid argues:

> understanding x and understanding that one understands x is one and the same epistemic state, of that inference; but a subject who grasps the inferential relationship that p stands to q, then precisely enjoys epistemic access to the inference. If she were unable to gain any access into the inferential relationship, then her epistemic state would not be one of grasping, but simply one of basing her belief that p on q (Janvid, 2018, p. 375).

The distinction here is important, especially in the case described above:

> I suggest that understanding requires the former kind of access, i.e., access to the ground, which in the case of understanding amounts to having access to the epistemic relation the relata stand to one another. It is precisely this relation that the subject grasps when she understands (Janvid, 2018, p. 375).

This is more straightforward in the context of proofs. Dumitru explores whether cognitive qualia play any part, even if not demonstrative, in understanding proofs:

> It occurs to me that it is quite reasonable to argue that "one knows that one understands that p" could be the case without necessarily being also the case that "one understands that one knows that p". Thus, I pretty much know that I understand stuff, but it is quite often the case that under the very same circumstances I don't understand that, and especially why, I know stuff (Dumitru, p. 183).

Think about scientific understanding as a spectrum, where very novel (paradigm shifting) discoveries are on the high end. The process is very creative, which burdens the agent with walking a very thin line between nonsense and truly novel, valuable outcomes. Multiple cognitive mechanisms and processes will be at play: *intuition*, inferring processes, grasp(ing), reflection (metacognitive processes), articulation, perhaps apperception, explanation, evaluation and so on. If the steps are not in sight, they're at some *underlying level*, then there's no real force to pull the demonstration forward. But perhaps there's some feeling to it. The feeling need not play any demonstrative role. One could think of it in terms of *fringes*, a phenomenal aspect of conceptual fluency, of rightness, meaning, where the reasoner is (...) *aware of relations and objects but dimly perceived* (James, 1980, Chapter IX, p. 258)[27] or in terms of qualia (epistemic, semantic, cognitive) (Dumitru, p. 181)

Lacking access to some inferences, the inferring process could continue (because it can use tacit grasps) to pull forward some inferences, until a block can surface and be grasped. So why do some grasps stay tacit? They are maybe left waiting for

[27] James, W. (1890). Chapter IX. The stream of thought. *The principles of psychology* (vol. 1) New York: Henry Holt & Co. p. 258.

some form that would fit a description and then an explanation. Perhaps one has the disposition to dismiss those tacit inferences based on previously acquired principles, scientific norms, or some other inferences that are glued together under the pressure of said principles. Perhaps some previous commitments offer epistemic stability that would contradict the commitments entailed by the inferences from the tacit grasps. Also, maybe that is why tacit grasps occur when the reasoner is caught off guard, not paying attention to the process, and these are therefore felt as "aha" moments. It's not that the content and attitude are separated, but rather, one is opaque to the other.

Think of long proofs and the steps taken when deriving one line from the other. Tacit graspings may allow for a high number of inferences (which humanly cannot be tracked) to be glued into a block, or a higher-level chunk. The chunk has more internal consistency than its component parts. When the chunk finds a thread (some scientific principle already available in one's scientific commitments) it can surface at a higher (explicit) level. At that moment, the shift in content (which must in the "aha" moment become explicit grasp) matches a shift in attitude (it is now *endorsable* because of a thread that shows some *hanging together*). But that need not be the case; the transition between tacit and explicit grasp need not be marked by a special surprising moment. "Aha" is not so much about surprise as about things fitting. Prior to an "aha" moment, there is a felt gap, *not-hanging-togetherness*, something off with a theorem, a dilemma, an argument, so on. Several criteria plausibly cluster for an "aha" moment. There need not be an *aha* moment – it is not a requirement for one to grasp. Instead of an "aha" moment, one could simply *see* the steps unfolding, recognize some sense of meaning or comprehension (a sort of cognitive or domain-specific qualia could be at play, not driving the inferences, but crawling through the grasp as it reaches a surface, *surveyable*[28] level).

The process of discovery is not to be seen as a performance. Interrogating an agent throughout this process about their scientific understanding in relation to the new discovery might prove disappointing at a functional level, if we are to take understanding as a performance.

The volume *Understanding and conscious experience* (Mărășoiu & Dumitru, 2024) is well-timed, published when the need to ponder the relation between the two is salient in relation to the "AI dilemma". If understanding is a peak performance, should we benchmark it and include LLMs in the scientific understanding measurement game? The functionalist path is alluring, but also risky. If they win the SU measurement game, we are faced with several choices: they understand, they don't

[28] I use the concept of surveyability ("überblickbar") as developed by Hilbert (*Zur Lösung dieser Probleme sehe ich keine andere Möglichkeit, als dass man den Aufbau der Zahlentheorie von Anfang an durchgeht und die Begriffsbildungen und Schlüsse in eine solche Fassung bringt, bei der von vornherein Paradoxien ausgeschlossen sind und das Verfahren der Beweisführung vollständig überblickbar wird*) – see Sieg (1999, p.30), and later by Wittgenstein as "übersichtlich" (*Wenn ich scrieb "der Beweis muss übersichtlich sein" so hiess das: Kausalitat spielt im Beweis keine Rolle. Oder auch: der Beweis muss sich durch blosses Kopieren reproduzieren lassen*) in RFM III (p.125, 47) Although it was translated by Anscombe as "perspicuous" – see RFM III, p.125e, §47. I'm using Felix Mühlhölzer's translation – see Mühlhölzer, F. (2006). "A mathematical proof must be surveyable" what Wittgenstein meant by this and what it implies. Grazer Philosophische Studien, 71(1), 57-86.

understand, they understand differently from us (different mechanisms are at play), they understand in some sense(s), but not in others, etc. Cognitive science should also have plenty to say, since it could help operationalize the term, providing explanations about understanding as a process (how it unfolds, how it arises, why it fails, how it fails). We should think about what kind of explanations we expect from cognitive science. Can cognitive science produce a scientific explanation of understanding? Do we expect a mechanistic one, a computational (representational) one, one based on dynamic systems, are we aiming for a Bayesian approach? Do we want *an integrative explanation* (one that should account for the cognitive, neural, phenomenological and behavioral aspect)? Should they also integrate at the levels or models (is it possible)? Do we need them to account for the phenomenal aspect of understanding? Do we settle for explanatory pluralism? Why go for a unifying theory when we could have fine-grained levels of analysis? The mechanistic approach could work out the architecture level, algorithmic approaches could work out the cognitive processes, and the phenomenological descriptions could provide some insight into what the experience might be like. Even though it might be tempting, some questions are left unanswered, such as the way in which all these fit together and which level is actually the relevant level for which sets of questions. Instead of clarifying, such explanations might lead to confusion.

AI systems and LLMs are an interesting tool, food for thought which might drive some necessary conceptual engineering, but a tool that also lessens the standards for what gets to count as a successful explanation. We must somehow protect explanations of understanding from being trivialized. Explanatory standards for what counts as understanding, in machines and humans, cannot be based on understanding as *performance*, or we might end up accepting thin criteria (it passes this benchmark, so it understands).

In sum, the book is an important contribution to epistemology of understanding, but not only, as its focus on the intertwinement of understanding and conscious experience may drive further research and shape new directions. As Dumitru already envisions in "Feeling the proof", new projects await fruitful developments, such as charting the diversity of cognitive qualia into semantic, geometric, logical, as well as another one of characterizing how these different kinds of qualia track specific abstract structures, *how the subjective aspects of our grasp anchor into the abstract realities thereby understood* (p. 191).

References

Alston, W. P. (1993). Epistemic desiderata. *Philosophy and Phenomenological Research*, 53(3), 527–551.

Barman, K. G., Caron, S., Claassen, T., & De Regt, H. (2024). Towards a benchmark for scientific understanding in humans and machines. *Minds and Machines*, *34*(1), 6.

Boghossian, P. (2014). What is inference?. *Philosophical studies*, 169(1), 1–18.

Chalmers, D. J. What we talk to when we talk to language models. https://philarchive.org/archive/CHAWWT-8

Coelho Mollo, D. (2021). Why go for a computation-based approach to cognitive representation. *Synthese*, 199(3), 6875–6895.

De Regt, H. W. (2017). *Understanding scientific understanding*. Oxford university press.

Di Paolo, E. A. (2018). The enactive conception of life. *The Oxford handbook of cognition: Embodied, embedded, enactive and extended*, 71–94.

Elgin, C. Z. (1996). *Considered Judgment*. Princeton: Princeton University Press.

Elgin, C.Z. (2022). Models as Felicitous Falsehoods. *Principia: An International Journal of Epistemology* 26 (1):7–23.

Froese, T. (2025). Sense-Making Reconsidered: Large Language Models and the Blind Spot of Embodied Cognition. https://doi.org/10.31234/osf.io/w7hg4_v2

Galli, G. (2024). Scientific Understanding and the Explanatory Integration in Cognitive Sciences. In: Aldini, A. (eds) Software Engineering and Formal Methods. SEFM 2023 Collocated Workshops. SEFM 2023. Lecture Notes in Computer Science, vol 14568. Springer, Cham. https://doi.org/10.1007/978-3-031-66021-4_7

Gardner, H. (1985). *The mind's new science: A history of cognitive revolution*. New York: Basic Books.

Graham, G., & Bechtel, W. (1998). *A companion to cognitive science*. Blackwell Publishing.

Haken, H., Kelso, J. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological cybernetics*, 51(5), 347–356.

Hohol, M. (2021). *Cognitive science: an interdisciplinary approach to mind and cognition*. In B. Brożek, M. Jakubiec, P. Urbańczyk (Eds.), *Perspectives on interdisciplinarity* (33–55), Krakow: Copernicus Center Press.

Hutto, D.D., Myin, E. (2014). Neural representations not needed - no more pleas, please. *Phenom Cogn Sci 13*, 241–256. https://doi.org/10.1007/s11097-013-9331-1

James, W. (1890). *Chapter IX. The stream of thought*. In *The principles of psychology* (vol. 1), pp.224-291. New York: Henry Holt & Co. http://dx.doi.org/10.1037/11059-000

Janvid, M. Getting a Grasp of the Grasping Involved in Understanding. *Acta Analytica* 33, 371–383 (2018). https://doi.org/10.1007/s12136-018-0348-5

Khalifa, K. (2012). Inaugurating understanding or repackaging explanation? *Philosophy of Science*, 79(1), 15–37.

Khalifa, K. (2016). Must understanding be coherent? In Grimm, S. R., Baumberger, C., and Ammon, S. (Eds.). (2016). *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science* (1st ed.). (pp.139–164). Routledge.

Khalifa, K., Islam, F., Gamboa, J. P., Wilkenfeld, D. A., & Kostić, D. (2022). Integrating philosophy of understanding with the cognitive sciences. *Frontiers in Systems Neuroscience*, 16, 764708.

Kirchhoff, M. D. (2025). Idealization and mental fictionalism. *Philosophical Psychology*, 1–22.

Mărășoiu, A. & Dumitru, M. (2024). *Understanding and conscious experience: philosophical and scientific perspectives*. New York & London: Routledge.

Menary, R. (2006). Attacking the bounds of cognition. *Philosophical psychology,19*(3), 329–344.

Mühlhölzer, F. (2006). "A mathematical proof must be surveyable" what Wittgenstein meant by this and what it implies. *Grazer Philosophische Studien*, 71(1), 57-86.

Noë, A. (2014). Concept pluralism, direct perception, and the fragility of presence. In *Open Mind*. Open MIND. Frankfurt am Main: MIND Group.

Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J., & Semenuks, A. (2019). What happened to cognitive science? *Nature human behaviour*, 3(8), 782–791.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). The Free Energy Principle in Mind, Brain and Behavior. In: *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press.

Reverberi, C., Pischedda, D., Burigo, M., & Cherubini, P. (2012). Deduction without awareness. *Acta psychologica*, 139(1), 244–253. https://doi.org/10.1016/j.actpsy.2011.09.011.

Sieg, W. (1999). Hilbert's programs: 1917–1922. *Bulletin of symbolic Logic*, 5(1), 1–44.

Strevens, M. (2024). Grasp and scientific understanding: a recognition account. *Philosophical Studies*, 181(4), 741–762.

Taylor, S. D. (2023). Afactivism about understanding cognition. *European Journal for Philosophy of Science, 13*(3), 43.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. MIT press.

Van Gelder, T. (1995). What might cognition be, if not computation?. *The journal of Philosophy, 92*(7), 345–381.

Wittgenstein, L., von Wright, G. H., Rhees, R., & Anscombe, G. E. M. (1978). *Remarks on the Foundations of Mathematics* (p. 19). Cambridge, MA: MIT press.

Zagzebski, L. (2001). Recovering understanding. In M. Steup (Ed.), *Knowledge, truth and duty* (pp. 235–251). New York: Oxford University Press.

Balkan Journal of Philosophy (2026), vol. 18, issue 1, 175 – 184

Book review

Vesselin Petrov (ed.). Process Philosophical Reflections on the Whiteheadian Intellectual Heritage. Cambridge Scholar Publishing, 2025, 275 p. ISBN: 978-1-0364-5663-4

[Reviewed by ***Lina Georgieva****]

Philosophical Reflections on the Whiteheadian Intellectual Heritage, edited by Vesselin Petrov, explores a variety of topics relevant to the current state of the world such as education, arts, ethics, community, medicine, psychology and artificial intelligence. It consists of fifteen chapters, carefully sorted in four thematic parts, all influenced by Alfred Whitehead's philosophy. This arrangement of thematic areas and the rich diversity of scientific and social fields is one of the volume's central concepts: that Whitehead's philosophy is not merely a historical system but a living framework capable of addressing contemporary problems.

Chapter One, "The Concepts of 'Creation' in the Late Philosophy of A. N. Whitehead", authored by Michel Weber, offers a philosophically ambitious reconstruction of the notion of creation in the late thought of Alfred North Whitehead, combining historical scholarship with metaphysical analysis. Weber's central aim is to show how Whitehead progressively replaces a traditional, theologically inflected concept of creation with the more radical and philosophically distinctive category of creativity, a shift that fundamentally transforms the ontological architecture of his system and redefines the relations between God, the world, and novelty. The chapter is structured around five stages in Whitehead's conceptual development, guiding the reader from early formulations through *Process and Reality* and into the late writings, thereby offering a coherent narrative of how creativity comes to assume the status of the Category of the Ultimate. One of Weber's contributions lies in his interpretation of creativity as an immanent ontological principle instantiated only in and through actual occasions, rather than as a transcendent cause standing behind the world. While Weber presupposes a high level of familiarity with Whitehead's corpus and technical vocabulary, this conceptual density is justified by the chapter's ambition, since it does not merely summarize Whitehead's views but seeks to resolve deep tensions within his system, particularly concerning the status of God and the grounding of genuine novelty. Overall, Chapter One constitutes a substantial and sophisticated contribution to Whiteheadian legacy, clarifying the philosophical stakes of Whitehead's transition from creation to creativity and providing a strong theoretical foundation for the rest of the volume.

*Institute of Philosophy and Sociology, Bulgarian Academy of Sciences; linageorgieva@ips.bas.bg

Chapter Two, "Critical Thinking: A Key to the Development of All Skills and Competencies in Life and Education" by Vesselin Petrov, advances a clear and programmatic thesis according to which critical thinking is not merely one skill among others but the foundational capacity upon which all meaningful competencies in education ultimately depend, a claim that is explicitly grounded in the educational philosophy of Alfred North Whitehead and situated within a broader process-oriented understanding of learning, development, and intellectual growth. Petrov opens his discussion with a historical overview of the concept of critical thinking, tracing its roots from classical philosophy through modern educational theory, a contextualization that functions not as a purely historical exercise but as a way of demonstrating that critical thinking has always governed how knowledge is acquired, evaluated, and applied, while at the same time providing an accessible yet non-superficial orientation for readers from both philosophical and educational backgrounds. The core theoretical contribution of the chapter lies in Petrov's systematic alignment of critical thinking with Whitehead's well-known stages of education – romance, precision, and generalization – through which he argues that critical thinking operates across all three phases of learning by animating curiosity in the stage of romance, ensuring rigor and discipline in the stage of precision, and enabling synthesis and the transfer of knowledge in the stage of generalization. Petrov further strengthens this argument by engaging contemporary literature on competencies and skills, particularly in the context of rapidly changing social and technological environments, and by adopting a distinctly normative position according to which critical thinking should not be treated as an optional enhancement or as a specialized subject confined to philosophy courses but should instead be integrated in all areas of education. By grounding contemporary debates about skills and competencies in Whitehead's process philosophy, Petrov offers a framework that is both philosophically and practically relevant, thereby serving as an effective bridge between classical process thought and current educational challenges.

Chapter Three, "Key Skills 2025 and Whitehead's Ideas on Education" by Denys Zhadiaiev, examines the relationship between contemporary educational discourse on future-oriented "key skills" and the enduring insights of Alfred North Whitehead's philosophy of education, advancing the central claim that, despite rapid technological and social transformations, the fundamental structures of human cognition and learning remain remarkably stable and therefore continue to sustain the normative and practical relevance of Whitehead's educational ideas. Through a comparative analysis of The Aims of Education and the "Key Skills 2025" framework promoted by institutions such as the World Economic Forum, Zhadiaiev argues that capacities now identified as crucial for the future – such as critical thinking, creativity, adaptability, problem-solving, and learning to learn – closely mirror those emphasized by Whitehead nearly a century ago, a convergence interpreted not as coincidence but as evidence that educational theory repeatedly rediscovers, under new terminologies, enduring philosophical insights about human learning. Resisting technological determinism, the chapter acknowledges the profound impact of digital technologies on everyday life while warning against the assumption that such changes fundamentally alter human cognitive architecture, and in doing so it implicitly reinforces Whitehead's

critique of "inert knowledge" by suggesting that future-oriented skills cannot be cultivated through automatic learning or purely technical training but instead require educational environments that foster curiosity, intellectual engagement, and integrative understanding. Although the chapter engages only modestly with empirical or institutional constraints, its primary contribution remains conceptual rather than programmatic, offering a philosophically grounded perspective that challenges superficial novelty and reaffirms the enduring value of process-based education, thereby complementing the surrounding contributions in the volume and strengthening its overall case for Whitehead's lasting intellectual legacy.

Chapter Four by Petya Klimentova offers a practice-oriented application of Alfred North Whitehead's philosophy of education to the field of media literacy, focusing on the Bulgarian educational context while drawing on broader European initiatives in order to address one of the most pressing challenges of contemporary education, namely how to cultivate critical and reflective media literacy in an era marked by information overload, disinformation, and algorithmic mediation. Grounding her analysis in concrete institutional and policy contexts through a secondary examination of the European Media Coach Initiative, Klimentova integrates empirical observation with philosophical reflection, using Whitehead's educational theory not as an abstract framework imposed from above but as a diagnostic and constructive tool for evaluating existing media literacy practices in Bulgarian secondary education. In doing so she is identifying significant gaps between policy aspirations and classroom realities. A central role in the chapter is played by Whitehead's three stages of education – romance, precision, and generalization –which Klimentova uses to argue that current media literacy initiatives tend to overemphasize technical skills and fact-checking techniques at the expense of curiosity, intrinsic motivation, and reflective synthesis, whereas a genuinely Whiteheadian approach would integrate all three stages in order to foster not only technical competence but also intellectual imagination and critical judgment. Emphasizing student-centered learning, active engagement, dialogue, and contextual understanding, the chapter aligns Whitehead's critique of "inert knowledge" with contemporary concerns about the passive consumption of digital content and reconceptualizes media literacy not merely as the recognition of misinformation but as the cultivation of habits of critical inquiry extending across disciplines and social contexts. By bridging philosophy, educational policy, and classroom practice, Klimentova demonstrates how Whiteheadian philosophy can inform educational reform in concrete national and institutional settings, making Chapter Four a thoughtful and practically grounded contribution to broader debates about education in digitally mediated societies.

Part Two of the book, devoted to The Future of Education, opens with Roland Cazalis's chapter "Whitehead and Intensification: The Power of Haptics in Early Education," which offers an innovative interdisciplinary exploration of early learning through the combined lenses of process philosophy, embodied cognition, and digital pedagogy. Inspired from the educational thought of Alfred North Whitehead, Cazalis advances the central claim that effective learning, particularly in early childhood, depends not only on cognitive abstraction but on the intensification of experience through bodily and sensory engagement, with special emphasis on haptics, a position developed within

a broader reflection on the societal impact of digital technologies and their transformative effects on learning environments. The chapter argues that the central pedagogical challenge today lies in integrating digital technology into a process-oriented framework that respects the holistic nature of human experience, maintaining that digital media should not replace embodied learning but should instead be designed to enhance and intensify it by anchoring abstract concepts in lived experience. A key theoretical contribution of the chapter is its reinterpretation of Whitehead's educational philosophy in light of contemporary insights from cognitive science and sensory studies, aligning Whitehead's emphasis on experience, feeling, and aesthetic intensity with recent research on haptic perception and suggesting that touch and bodily interaction play a crucial role in meaning-making and cognitive development. Cazalis further critiques reductionist educational models that treat learning as the acquisition of discrete skills or competencies, arguing instead for a comprehensive pedagogical framework attentive to the learner, a position that resonates strongly with Whitehead's critique of inert knowledge. By creatively extending Whiteheadian process philosophy into the domain of haptic learning and digital pedagogy, the chapter opens new avenues for thinking about how technology, sensation, and education can be meaningfully integrated, making it of particular interest to scholars working at the intersection of philosophy of education, cognitive science, and educational technology.

Chapter Six "The Aims of Education for the Future and Whitehead's Philosophy" by Štefan Zolcer offers a reflection on the "Aims of education" in the context of accelerating social, technological, and ecological change. The chapter challenges conventional models of education that prioritize adaptation to existing or past social conditions and argues instead for an education oriented toward shaping a sustainable future which is in line with the educational ideas of Alfred North Whitehead. Zolcer proposes that education must cultivate capacities that enable individuals and societies to respond creatively and ethically to unforeseen futures. A central conceptual contribution of the chapter is the introduction of a three-level model for analyzing educational problems, which Zolcer suggests is partly applicable to broader global crises. Although the model is not developed as a rigid framework, it serves as a heuristic for distinguishing between individual, collective, and global dimensions of educational aims. This multi-level perspective allows the author to move beyond purely individualistic or technocratic conceptions of education and to situate learning within wider moral, social, and ecological contexts. Qualities such as critical thinking, empathy, flexibility, cooperation, emotional stability, and environmental awareness are presented as central educational aims. One of the chapter's strengths is its explicit engagement with contemporary crises, including moral disorientation, environmental degradation, and the erosion of social cohesion. Zolcer convincingly argues that these challenges cannot be addressed through technical expertise alone. Instead, they require an education that fosters holistic human development and a sense of responsibility toward others and the planet. The chapter will be of particular interest to those concerned with educational philosophy, sustainability, and the moral purposes of schooling.

Chapter Seven "The Future of Education: The Philosopher's Responsibility for Education in the Age of AI" by Vesselin Petrov addresses one of the most urgent

questions facing contemporary education: how education should respond to the rapid development of artificial intelligence. The chapter advances an argument about the responsibility of philosophy in shaping educational futures rather than merely interpreting past traditions. Rather than treating AI as a neutral tool or an external disruption, Petrov frames it as a transformative tool that challenges foundational assumptions about knowledge, teaching, agency, and intelligence. The author argues that weak AI – understood as advanced but non-autonomous systems – can play a constructive role in education by supporting learning processes and reshaping the teacher's role. In this scenario, AI functions as an aid or partner rather than a replacement for human educators. Strong AI, by contrast, raises far more radical questions. If artificial systems were to achieve a level of intelligence comparable to human intelligence, Petrov suggests, they would demand ethical and philosophical consideration akin to that afforded to extraterrestrial intelligence. This analogy between strong AI and extraterrestrial intelligence is one of the chapter's most provocative elements. It serves to underscore the epistemic and moral uncertainty surrounding advanced AI and to challenge anthropocentric assumptions about intelligence and education. From a Whiteheadian perspective, this move is consistent with a process ontology that resists fixed hierarchies and emphasizes relational becoming. The chapter strongly emphasizes the continuing importance of the human teacher. Even as AI systems become more capable, Petrov argues that education must remain a fundamentally human process grounded in value, judgment, and responsibility. This position resonates with Whitehead's critique of inert knowledge and reinforces the idea that education is not merely about information transfer. While the chapter is deliberately speculative with no concrete examples of current AI-in-education practices, it offers a bold reflection on education in the age of artificial intelligence.

"The Value of Philosophical Approaches as Art Forms: Another Sarcasm or Consecutive Analysis?" by Denys Zhadiaiev is the opening chapter for Part Three: Arts and Ethics. It offers a critical and reflective examination of the relationship between philosophy and art, questioning contemporary tendencies to blur or dissolve the boundaries between philosophical argumentation and artistic or literary expression. While grounded in the broader context of process philosophy, the chapter engages more directly with classical epistemological and logical concerns than with strictly Whiteheadian metaphysics. The central problem addressed in the chapter is whether philosophy can – or should – be understood as a form of art. Zhadiaiev approaches this question with a critical stance toward postmodern discourses that treat philosophical writing primarily as expressive, rhetorical, or creative performance. Rather than dismissing the aesthetic dimension of philosophy outright, the author seeks to clarify the criteria by which philosophical discourse can be distinguished from non-philosophical forms of expression. Drawing on Aristotelian logic and Kantian reflections on philosophical method, Zhadiaiev argues that philosophy, unlike art, is constrained by rational coherence and argumentative responsibility. While artistic expression may thrive on ambiguity, contradiction, and affective resonance, philosophy must remain accountable to conceptual clarity and logical consistency. The chapter also addresses the contemporary cultural context in which philosophical discourse increasingly overlaps with political, social, and ideological narratives. Zhadiaiev critically examines claims that

various forms of social or political commentary should be recognized as philosophy simply by virtue of their thematic ambition or rhetorical sophistication. Against this trend, he insists that philosophy requires a stable categorical framework and a commitment to rational justification. By reaffirming the importance of logical consistency and conceptual discipline, Zhadiaiev contributes a clarifying voice to ongoing debates about the nature and limits of philosophical practice. The chapter complements the volume's broader engagement with creativity and process by insisting that creativity in philosophy must operate within, rather than against, the demands of rational analysis.

Chapter Nine "Process Architecture: An Ecological Alternative to Modern Architecture" by Maria-Teresa Teixeira extends process philosophy into the domain of architectural theory, offering a philosophically ambitious critique of modern architecture and its ecological consequences. Drawing explicitly on the process metaphysics of Alfred North Whitehead, the chapter argues that prevailing architectural paradigms are rooted in a static, mechanistic worldview that has contributed significantly to environmental degradation. In response, Teixeira proposes "process architecture" as an ecologically grounded alternative. Teixeira convincingly frames modern architecture as complicit in an extractive economic model that prioritizes efficiency, abstraction, and formal novelty over lived experience and ecological integration. This critique resonates with Whitehead's broader rejection of substance-based metaphysics and his insistence on understanding reality as dynamic, relational, and processual. A central theoretical point of the chapter is the architectural philosophy of Christopher Alexander, particularly as developed in *The Order of Nature*. Teixeira presents Alexander's notion of wholeness as a concrete architectural analogue to Whitehead's concept of organic unity. Buildings, on this view, are not isolated objects but living participants in a broader field of relations involving human activity, social practices, and natural environments. By drawing this parallel, the chapter effectively bridges metaphysical theory and architectural practice. Teixeira argues that process architecture seeks to restore continuity between built environments and natural processes, thereby counteracting the alienation produced by modernist design philosophies. Rather than treating Whiteheadian philosophy as a metaphorical backdrop, Teixeira integrates process concepts directly into architectural analysis. The chapter gestures toward practical implications, suggesting that alternative architectural practices inspired by process thought have already been explored and could be further developed.

Chapter Ten "Aesthetic Experience and Normativity in a Process Axiological Perspective" by Sylvia Borissova offers a philosophical exploration of aesthetic experience and normativity through the lens of process axiology. The chapter situates value not as a static property or external standard but as an emergent feature of ongoing processes of becoming. The argument opens with a brief historical overview of axiology, outlining its emergence as a distinct philosophical field at the turn of the nineteenth and twentieth centuries. This historical framing clarifies the conceptual stakes of Borissova's project: to rethink value beyond both subjectivist relativism and objectivism. By situating Whitehead's philosophy within this tradition, the chapter highlights the originality of his approach to value linked to creativity, feeling, and experience. Borissova's central argument is that, within a process framework, aesthetic experience is not a marginal or purely subjective phenomenon but a fundamental

mode of value realization. The author emphasizes that, in Whiteheadian axiology, the products of creativity are not limited to works of art. All "concrescences" of actual entities – whether organic or inorganic – exhibit aesthetic dimensions insofar as they achieve intensity, harmony, and balance. This move effectively dissolves the sharp boundary between aesthetics and metaphysics. The chapter also addresses the relationship between aesthetics and normativity. Borissova argues that norms arise from patterns of value that are repeatedly realized and recognized within experiential processes. This account avoids reducing normativity to mere convention while also resisting transcendental moralism. In this respect, the chapter complements other contributions in the volume that emphasize the ethical implications of Whitehead's thought, while maintaining a distinct focus on aesthetic experience as the primary locus of value. By articulating a process-oriented account of aesthetic experience, Borissova demonstrates how Whitehead's philosophy can offer a coherent and non-reductive foundation for axiology. The chapter deepens the volume's engagement with questions of value and meaning.

Chapter Eleven "The Ethical Relevance of Whitehead's Thought" by Federico Giorgi is the closing text of Part Three: Arts and Ethics. The chapter's central aim is to demonstrate that Whitehead's metaphysics, often read primarily as a cosmological or ontological system, contains a distinctive ethical orientation grounded in value, feeling, and creative advance. Giorgi's contribution is notable for its conceptual clarity and its effort to systematize Whitehead's ethical insights without distorting their process foundations. The chapter is organized into several interconnected sections that guide the reader from metaphysical premises to ethical implications. Giorgi begins by outlining the basic features of Whitehead's organicist metaphysics, emphasizing the primacy of process, relationality, and becoming. This groundwork is essential for the ethical argument that follows, as it establishes that morality, in a Whiteheadian framework, cannot be understood as obedience to fixed rules or external norms but must instead be rooted in experiential value and creative transformation.The analysis of key ethical notions such as importance, value, and beauty enables Giorgi to argue that these concepts are not peripheral but foundational to Whitehead's moral vision.

The discussion of moral feelings is another significant contribution to the volume. Giorgi interprets moral feelings as emergent phenomena arising from the relational structure of experience rather than as subjective emotions detached from objective reality. In its latter sections, the chapter extends Whiteheadian ethics beyond human-centered concerns by addressing issues related to space exploration and cosmocentric ethics. This extension is particularly compelling, as it demonstrates the flexibility and contemporary relevance of Whitehead's ethical framework. By demonstrating how moral value, feeling, and responsibility emerge from a process metaphysics of creativity, Giorgi offers a compelling alternative to both rule-based and purely subjective ethical theories.

Chapter Twelve "Re-Thinking Community and Communication from a Processual Perspective" opens Part Four of the volume, which turns from education, arts, and ethics to a set of interdisciplinary reflections on community, medicine, psychology, and artificial intelligence. As the first contribution in this final section, Maria Regina

Brioschi's chapter establishes a conceptual framework that foregrounds relationality and communication as central concerns of contemporary process philosophy. The main argument of the author is that process approach is particularly well suited to addressing the social and communicative challenges of late-modern societies. Brioschi contends that traditional, substance-based models of community and communication are increasingly inadequate for understanding social life in a world characterized by fluid identities, digital mediation, and complex networks of interaction. In response, she proposes a process-oriented framework that emphasizes becoming, relationality, and experiential continuity over static notions of social order. Rather than treating community as a fixed entity defined by shared attributes or institutional boundaries, Brioschi presents it as an emergent process constituted through ongoing patterns of interaction and mutual influence. The discussion of communication builds upon this processual understanding of community. Brioschi argues that communication should not be reduced to the transmission of information between pre-existing subjects. Instead, it is a constitutive activity through which subjects and social relations are continuously formed and transformed. One of the most distinctive aspects of the chapter is its engagement with Whitehead's concept of evidence. Brioschi introduces evidence not merely as epistemic justification but as a lived, experiential phenomenon that emerges within communicative processes. This move allows the author to bridge epistemology and social theory, suggesting that shared understanding and communal coherence arise from patterns of experiential confirmation rather than from abstract norms alone.

Chapter Thirteen "Processual and Structural Approaches to End-of-Life Narratives Analysis" continues Part Four of the volume, which addresses issues at the intersection of philosophy, medicine, psychology, and social life. Anastasiia Zinevych's contribution offers an end-of-life narrative, bringing process philosophy into dialogue with narrative analysis in clinical and existential contexts. The chapter's central aim is to compare and critically assess two hermeneutical strategies for analyzing end-of-life narratives: a processual (synthetic) approach and a structural (analytic) approach. Zinevych does not present these strategies as mutually exclusive alternatives but rather as complementary levels of interpretation, each capturing different aspects of narrative meaning. The processual approach seeks to grasp the narrative as a lived whole, emphasizing fluidity, temporality, and the unfolding of experience beyond discrete narrative elements. This strategy aligns closely with process philosophy's emphasis on becoming, continuity, and experiential depth. By contrast, the structural approach focuses on identifying stable narrative components – such as images, motifs, plots, and macroplots – that organize and give coherence to the narrative. Zinevych demonstrates that, even within highly fluid and emotionally charged end-of-life narratives, certain structural patterns remain remarkably persistent. The author argues that narrative content and narrative structure operate on different ontological levels. While the surface content of end-of-life narratives may shift in response to changing circumstances, emotions, and physical conditions, the underlying meta-plot often remains stable. The chapter also raises an important and unresolved question concerning motivation. Zinevych notes the difficulty of establishing a precise methodological link

between narrative motifs and the psychological category of motivation. While motifs and plots can be identified with relative clarity, the motivational forces that shape narrative selection and emphasis remain elusive. Rather than presenting this as a shortcoming, the author frames it as a productive problem that points toward future interdisciplinary research combining philosophy, psychology, and narrative medicine.

Chapter Fourteen "Whitehead's Concept of Symbolic Reference: Insights into Autism" brings process philosophy into dialogue with contemporary discussions in psychology and autism studies. In this chapter, Elisabetta Angela Rizzo offers an original and conceptually sensitive application of Alfred North Whitehead's theory of perception – specifically the notion of symbolic reference – to the understanding of autism. The chapter's central contribution lies in reframing autism not primarily as a deficit in social communication, but as a difference in perceptual and symbolic organization. Rizzo grounds her analysis in Whitehead's distinction between two fundamental modes of perception: causal efficacy and presentational immediacy. Symbolic reference, on Whitehead's account, emerges from the integration of these two modes, allowing perceptual experience to be stabilized, abstracted, and linguistically articulated. The chapter argues that autistic perception often involves a different balance between these modes, leading to distinctive sensory experiences that shape communication, language use, and social interaction. One of the chapter's contributions is its shift away from deficit-based models of autism. Rizzo shows how variations in perceptual processing can influence the formation and use of symbols, thereby affecting linguistic expression and interpretation. This approach challenges simplified accounts of autism that focus narrowly on social cognition, suggesting instead that communicative misunderstandings often arise from deeper perceptual divergences between autistic and non-autistic individuals. Whitehead's concepts are not employed metaphorically but are carefully adapted to support empirical findings from autism research. This careful integration allows the chapter to bridge philosophy, psychology, and neurodiversity studies.

Chapter Fifteen "A Whiteheadian Approach to the Perception of Molecular Machines" concludes Part Four of the volume with an interdisciplinary contribution that extends Whiteheadian process philosophy into the realm of molecular biology and nanoscience. In this chapter, Roland Cazalis proposes an original interpretation of molecular machines – most notably viruses – through the perceptual and ontological framework of Alfred North Whitehead. The chapter's central ambition is to demonstrate that Whitehead's concepts of prehension, concrescence, and satisfaction can clarify the logic underlying interactions between inert matter and living systems. Cazalis begins by challenging conventional, anthropocentric notions of perception. He adopts Whitehead's expansive view according to which perception is a fundamental feature of reality itself. At the most basic level, perception consists of blind, non-conscious acts of prehension through which entities respond to and incorporate aspects of their environment. This reconceptualization allows Cazalis to analyze molecular machines not merely as mechanical devices but as participants in processual relations of sensing and response. A central topic of the chapter is the case of viruses, which occupy an ambiguous position between inert matter and living organisms. Cazalis interprets viral activity – particularly the integration of viral genomes

into host cellular structures – as a process of concrescence culminating in satisfaction. On this account, the successful incorporation of viral material into a cellular system represents the completion of a processual trajectory guided by value and historical relatedness. This interpretation reframes viral behavior as neither purely mechanical nor purposive in a conscious sense, but as governed by an organic logic embedded in the structure of reality. By interpreting molecular machines through a processual lens, Cazalis suggests that the distinction between inert and living matter is better understood as a gradient of organizational complexity rather than a categorical divide. This view aligns closely with Whitehead's panexperientialist commitments and opens new avenues for philosophical engagement with contemporary science.

In conclusion, *Process Philosophical Reflections on the Whiteheadian Intellectual Heritage* offers an interesting view of the contemporary relevance of Alfred North Whitehead's philosophy across a wide range of theoretical and practical areas, including education, ethics, aesthetics, community, psychology, artificial intelligence, and the philosophy of science, thereby showing how process thought can meaningfully address some of the most pressing problems of the present. The authors demonstrate deep knowledge and understanding of Whiteheadian concepts making a valuable contribution to contemporary philosophy and also broader debates in philosophy of education, reaffirming process thought as an effective instrument for working in an interdisciplinary field. The overall feeling from reading the collection is of a philosophically rich encounter with a living tradition of thought that continues to generate original insights into contemporary educational, ethical, technological, and cultural problems.