Contents

Research Articles

Book Review

Organon F 33 (1) 2026: 2–32
https://doi.org/10.31577/orgf.2026.33101

ISSN 2585-7150 (online)

RESEARCH ARTICLE

Dispositional Universals

Tomáš Károly*

Received: 10 October 2023 / Revised: 2 October 2025 / Accepted: 6 December 2025

Abstract: The ontology developed in this article is a version of the bundle theory of universals. These universals are dispositional in nature. To possess a powerful character and to be empirically existent, these universals require a material substratum. This substratum is conceived as physical ether or Plato's chora, in which universals are like waves in the sea. The world is composed of basic dispositional universals, such as charge, mass, and spin. These basic dispositional universals bundle together to form elementary particles, which in turn compose more complex structures. An object is thus a composition of basic dispositional universals, whose sustained unity is explained by their mutual dispositional manifestations. The assembled whole already exhibits different manifestations than its parts and therefore constitutes a new kind of dispositional universal of a higher order. Hence, a graded hierarchy is present in the structure of dispositional universals. The article defends the view that dispositions are universals rather than tropes, since what we observe in the physical world are identical, numerically repeatable properties, not merely similar particulars. The proposed theory of dispositional universals stands in opposition to Humean conceptions, as the thesis that "whatever is conceivable is metaphysically possible" does not hold. The lawfulness of the world is conditioned by the dispositional nature

* University of Ss. Cyril and Methodius in Trnava
https://orcid.org/0000-0002-4344-8556
✎ Department of Philosophy and Applied Philosophy, Faculty of Arts, University of Ss. Cyril and Methodius in Trnava, Námestie Jozefa Herdu 577/2, 917 01 Trnava, Slovak Republic
✉ tomas.karoly@ucm.sk

of universals. Scientific models that aim to explain and predict events should therefore be adapted to this metaphysical framework.

Keywords: Bundle theory; dispositions; kinds; laws of nature; universals.

1. Introduction

The world is made of stable objects; our bodies are held together; the dog in the garden is always barking; the neighbour's donkey has not begun smoking a pipe and talking about definite descriptions. Whether inside or outside our dwelling, we recognise objects, name them and know how these objects are manifested. We instinctively avoid a moving truck, and a dog will avoid it, too. Not only humans but animals, too, have the capability of predicting the behaviour of objects known to them when they are in contact with such and such circumstances or exhibit certain known states. The whole world seems to be ruled by order in the form of immutable laws. "Laws are supposed to be somehow the things that activate the world: the things that add the necessity and possibility to it and thereby make events happen" (Mumford 2004, 14). The world is made up of objects, and we classify these objects into kinds based on their common properties and give them names: dog, human, rock... The world is also governed by laws that we can also classify into kinds: laws of motion, laws of optics, laws of common sense... For example, Newton's law: $F=ma$, Snell's law: $\sin\theta_1 n_1 = \sin\theta_2 n_2$, common sense law: a stove burner that's glowing red is hot and will cause burns if touched. These laws apply equally throughout the world. My aim will be to explain how kinds of objects originate and what is the source of regularity in the world.

In this text, I argue that the world is composed of basic universals that are dispositional in nature; I refer to these as dispositional universals. The instantiation of universals as materially grounded entities presupposes the existence of a substratum, which I designate as ether, through which their dispositional capacities can manifest. The manifestation of these basic dispositional universals allows them to compose into stable objects. The stability of such bundles is made possible by the ethereal substratum in which the dispositional universals are instantiated.

Basic dispositional universals are structured into higher-level wholes, and their manifestation takes the form of a stable particular that belongs to a specific natural kind. The dispositional nature of these kinds of objects necessarily forces them to always manifest themselves under given circumstances in the same way, and to do so everywhere in the universe, at all times, and in all possible worlds in which these kinds would occur.

2. Laws of Nature

A central goal for science is to produce explanations from which it is capable of generating testable predictions (Douglas 2009, 445). One of the most basic models used to explain an event is the deductive-nomological (D-N) model of Hempel and Oppenheim (1948). The explanation takes the form of a deductive judgment: in the explanans are statements of antecedent conditions, $C_1, C_2, \ldots, C_k$ together with general laws $L_1, L_2, \ldots, L_k$, and in the explanandum are descriptions of the empirical phenomenon to be explained E (Hempel and Oppenheim 1948, 138). If we know the premises (explanans), we are also able to deductively predict the empirical phenomenon in the form of a statement in conclusion (explanandum). The prediction is also a test of the truth of the statements contained in the premises of this model.

A law of nature can be expressed as a conditional "If *F* then *G*," or more specifically "Whenever *F*, then *G*," or "All *Fs* are *Gs*." For example, "If the temperature of a gas increases while keeping the pressure constant, its volume will also increase"; "Whenever a current flows through a conductor, a magnetic field is produced around it" or "All metals are electrically conductive."

It is necessary to note a point of interest, that mental ability, such as logical reasoning, serves us in D-N model for explaining and also predicting natural events. More than one philosophical dispute on the nature of the necessity of the world arose specifically due to this mixing of the *logical* with *natural* necessity. We can imagine that anything logically possible may be contained in the premises, and the form of judgment will guarantee correct logical deduction from them. By this mixing, physically unrelated facts can also appear in the D-N model: From the length of the shadow cast by the flagpole and the distance from the end of the shadow to the top of the flagpole, we will explain *why* the flagpole is so high using the Pythagorean theorem. This is a mistake, because the shadow is not the cause of the size of the flagpole. For other examples, see (Bromberger, 1966).

In order to fix the strength of the connection between F and G and to avoid mere accidental generalisations, such as “All solid spheres of gold have a radius less than 1 km,” a counterfactual dependency must exist between them. As Lewis (2001, 2) puts it: “If it were the case that F, then it would be the case that G.” For a counterfactual statement in the laws of nature to be true, it should be connected to an empirical necessity. Laws should “express some sort of necessary connection between their antecedents and their consequents that is missing between the antecedent and the consequent of true but merely accidental generalisations” (Rosenberg and McShea 2008, 44). The cause F and the effect G should be ontologically necessarily connected.

> When we say that there is a causal connection between two consecutive events, we mean that there is some kind of law connecting them, the earlier event being called the cause, and the later the effect. The question then arises as to what is the specific nature of the nexus between them. Is there any criterion permitting us to say that a given natural event is the effect of another? This question is as old as natural science itself... (Planck 1963, 44)

Philosophy offers several interpretations of the nature of nexus, as mentioned by physicist Planck. First, I will briefly introduce Humean theories. I will devote only minimal space to them, since they have already been extensively discussed in philosophical literature. I will then reject these theories, bypass many others, and focus exclusively on a specific non-humean theory of dispositional universals.

According to strict followers of Hume’s philosophy, the laws of nature supervene on constant conjunctions of events. In this view, the world is not governed by laws; the fact that we expect an effect to follow a cause is merely a habit of thought, not a manifestation of any natural necessity observable in the world. Since we do not perceive any necessary connection in nature, Hume concludes that cause and effect are distinct and separate events. The fact that specific causes have been followed by specific effects does not entail that this must always be the case in the future (Hume 1960, Part III, Section XIV). “Statements of laws, then, are merely descriptions of the most significant regularities that happen to occur" (Hildebrand 2023, 2). This Humean model is, to me, deeply unintelligible. That metals have always expanded when heated, that stones have fallen from heights, that dogs have typically barked, that human beings have philosophised throughout human history, that the Taj Mahal has held together in the same place–are all these regularities and persistent stability merely the products of accident? I would be afraid to lie down in my bed, for it might turn into an abyss at any moment.

Humean supervenience "is the doctrine that all there is to the world is a vast mosaic of local matters of particular facts, just one little thing and then another" (Lewis 1986a, ix). According to Lewis's modal realism, possible worlds with their own Humean regularities over which the laws of nature supervene do physically exist. In his view, two theses apply: (1) absolutely every way that a world could possibly be is a way that some world is, and (2) absolutely ever way that a part of a world could possibly be is a way that some part of some world is (Lewis 1986b, 86). This conception acknowledges that there are also worlds in which metals do not conduct electricity, salts do not dissolve in water and jellyfish in the sea have DNA like B. Russell had in our world. As for our current world, well Lewis (2001, 73) restates Ramsey's theory of lawhood: a contingent generalization is a law of nature if and only if it appears as a theorem (or axiom) in each of the true deductive systems that achieves a best combination of simplicity and strength.

In these Lewisian-Humean worlds, no ontologically separate laws of nature or necessary connections exist; the world is contingently regular for some mysterious reason; nothing supports this regularity, and therefore it makes no sense to speak of some natural necessity. An account of what grounds regularity, in the form of laws of nature, appears in the influential theory of Dretske (1977), Tooley (1977) and Armstrong (1989). A law of nature is second-order universal N that connects first-order universals, F and G. This relation is denoted as N(F,G). According to Armstrong (1985, 85) the relation $N(F,G) \rightarrow \forall x\ (Fx \rightarrow Gx)$ applies, which means that Humean regularity is derived from the law of nature, not the other way around. "This non-logical necessitation [N] entails a constant conjunction at the level of first-order particulars (with reservations still to be made), but the constant conjunction does not entail the necessitation" (Armstrong 1978, 90).

Here, the law of nature N, which necessitates the regularity of world, is dominant. According to Armstrong's combinatorial theory of possibility, as many possible worlds are conceivable as there are conceivable combinations of universals among themselves and thus also combinations of various laws of nature. For this reason, Armstrong talks about the so-called weak necessity, because the content of this necessity differs from world to world, depending on the laws instantiated there. Armstrong is an advocate of universals *in re*, and these need not be instantiated now because the past, present, future are equally real (Armstrong 1985, 82).

Armstrong's universals are inactive categorical properties that are connected through contingent necessity. Therefore, possible worlds in which donkeys talk exist, because the law of nature allows them to. It is possible to combine the universal of donkeyness with some universal including the possibility of talking, and thereby we get the possibility of a talking donkey. D. M. Armstrong says that here is a strong reason to think that a talking donkey is possible (Armstrong 1989, 101–102). A glass vase is breakable because certain circumstance obtain the categorical arrangement of the glass F together with the law of nature N will allow the vase to break G. The general problem of this theory is how N is able to act on the universals F and G. How can something act or connect when all properties are passive and powerless?

Even if the laws of nature were some Platonic non-spacetime entities, it is very problematic to explain how it is possible that they can act on passive matter and dictate what it should do (cf. Mumford 2004). The same problem arose in Descartes's dualism, which the 17th century occasionalists explained by means of divine action. God desires (F,G), therefore (F,G) will happen. The world's existence is dependent on God's will. God ceases to will and the world ceases to exist; no bodies have the power in themselves to move by themselves; God alone is the essential cause of their motion (Malebranche 1997, Book Six). In the occasionalist world, God somehow miraculously moves passive objects or re-creates the world (Ott 2009, 71) based on his own will. He is the cause of objects behaving lawfully, and he can also decide to perform a miracle and violate customary regularities of the world (e.g. the case of the Eucharist).

All these scenarios of the world, in which the laws of nature somehow enter into world events, are very complicated, inconceivable. For example, every time an electron meets a positron, the invisible delicate fingers of God appear in the form of the law of nature, and they send these two elementary particles into a state of annihilation. God's hands constantly hold us to the Earth so that we do not fly off into space. These divine hands create Descartes (1996, 33–34) anew with every moment. Perhaps, this divine power stood before the creation of one of the best possible worlds, and the regularity is formed of pre-established harmony; the world is set up as a great mechanism with inactive monads (Leibniz 1985; 1989). Or, as I will further argue in this text, this power is contained within the objects themselves, or more precisely, objects are shaped by force properties–dispositional universals. These universals are of a modal character, and the entire universe may be reflected in them as in monads; but unlike Leibniz's monads, these are causally active thanks to ethereal physical substratum.

3. General Characteristics of Disposition Properties

The word disposition has a Latin origin *dispositio* and its Greek equivalent is *diathesis*. Disposition means something like "orderly arrangement" (Jansen 2009, 24). We can also think of disposition as dynamis (Aristotle's term), power (Locke's term), ability, potency, capability, tendency, potentiality, proclivity, capacity and so forth (Choi and Fara, 2021). Dispositions have a *directedness*, that is a power for, or to, some outcome (Molnar 2006, 57).

A disposition is some internal property of an object which is manifested under a given stimulus, always necessarily in the same way in all possible worlds. Always when we immerse salt in water with *ceteris paribus* conditions (no blockers such as finks[1] or antidotes[2] are present), the salt dissolves. It would dissolve wherever the same or similar conditions were present, because dissolving in water under the given conditions is an essential property of salt. Bird (2007, 60) tells us we can capture the definition of a disposition's manifestation through the conditional: $\forall$x (*ceteris paribus* ((Dx & Sx) $\rightarrow$ Mx)), where D is the disposition, S is a stimulus and M the manifestation.

[1] The term was coined by C. B. Martin (1994).

[2] The term was coined by A. Bird (1998).

If we were to agree with Humean metaphysics, then it is not necessary truth that salt will dissolve under the given conditions, because it is conceivable that it would not dissolve. In the context of dispositional theories, however, it is true that every object which has basic chemical structure NaCl must inevitably dissolve under the given conditions; if this were not the case, then we do not have an NaCl object in front of us, but something else similar to it. Likewise, if something is H_2O it must also manifest itself as H_2O; even if there were a liquid XYZ (Putnam 1973) very similar to water on Earth, it could no longer be water as we understand it, but another substance. According to B. Ellis (2001, 48), in scientific essentialism there are genuine causal powers, capacities and propensities which exist in nature as universals, and therefore they are the same in all possible worlds. In my concept of dispositional universals, H_2O should be a complex universal that, when instantiated in any possible world, must always manifest as being watery under the same conditions. H_2O in one world and H_2O in another world are numerically different but qualitatively identical. H_2O and XYZ are both numerically and qualitatively different; therefore, the identity relation cannot apply between them.

4. Tropes or Universals

The medieval dispute between nominalism and realism endures even into the present. The importance of this topic lies in the fact that if universals do not exist, then our descriptions of nature and our generalizations are mere fiction. Knowledge of nature cannot exist, because in a world where no two properties are identical, our employed vocabularies and laws cannot refer to individual entities.

Proponents of the existence of dispositional properties or powers are also nominalists (Heil 2003), (Molnar 2006), (Martin 2008), and realists (Mumford 1998; 2004), (Ellis 2001), (Bird 2007), (Tugby 2013). I advocate the position of the bundle theory of dispositional universalism. My version, however, is a moderate form of the bundle theory of universals, because for universals to become materialized and active, they must have a substrate – some kind of sea in which waves are realized, like Plato's *chora* (χώϱα) in the *Timaeus*, or an ether as in physics.

Proponents of tropes say there are such things as attributes, but they deny that attributes are multiply exemplifiable entities. According to this form of nominalism, it is impossible for numerically distinct things to have numerically one and the same attribute. No two particulars share a single attribute. There is always a difference between numerically different particulars, however slight, in shade of colour, in shape, in size, in weight, and so on (Loux 2006, 72). Particulars, such as tropes, unlike universals, can never be qualitatively identical but only similar. "Tropes are *particularized properties*" (Ehring 1997, 11). According to Ehring, tropes are not necessarily momentary; there are also persisting tropes. The redness of the apple (endurantistically) persists over time. Not all tropes persist in time, however; some are momentary entities, like the blue of a movie-image sky (Ehring 1997, 14). The position of nominalists is rather strong in the case of complex macro-objects. When we look at one Chihuahua and another, we see the likenesses, not the sameness. Chihuahuas can vary in the arrangement of their fur, their shade, their character, etc. The word "Chihuahua" refers only to particulars whose characteristics they resemble. Today, biologists particularly struggle with defining species and determining whether they exist as real ontological entities.

Universals usually mean *properties, kinds, relations.* They are repeatable entities. Table salt *here* and table salt *there* share an identical chemical structure, colour and taste; they fall into a kind of table salt, and when they are in relation with water under the same conditions, *ceteris paribus*, they manifest identically by dissolving. According to Plato's realism *ante re*, not all universals must be instantiated somewhere; universals are non-spatio-temporal entities, while Aristotle's realism *in re* only admits the existence of instantiated universals. Plato embraces a "two-worlds" ontology; his "universals are ontological 'free floaters' with existence conditions that are independent of the concrete worlds of space and time" (Loux 2006, 41).

The position of the bundle of universals seems to be ontologically economical, without the need for an ontologically obscure substance as the bearer of properties. Objects without substance are bundles of universals. Objects are identical to bundles of *immanent universals* (Lafrance 2015, 202). According to Lafrance, for universals to be immanent, they need to be in fact instantiated by a region of space (Lafrance 2015, 203). "Universals distributed over space are grouped, or bundled up, together. ...A blue cube is identical to the bundle containing *blueness*, *cubic* and some other universals of *mass*, of *rigidity*, etc" (Lafrance 2015, 203–204). In the bundle theory in general, the problem is to explain how the "co-instantiation" of universals holds the bundle together. Some also call this relationship of the relation of the bundle "compresence," "collocation," "combination," "consubstantiation" and "coactuality" (Loux 2006, 91); this relation is generally considered to be primitive and undefinable. It is also a problem to explain why incompatible universals, such as squareness and roundness, do not occur in a mutual bundle (Lafrance 2015, 213–214). If universals should be powerless, then it should be possible to think about the "fusion" of everything with everything else. My proposal of dispositional universals can resolve these mentioned problems. Before we present them in more detail, I would still answer the question why dispositional universals and not dispositional tropes.

In the case of the existence of macro-objects, it is obviously less problematic to claim that the world is made up of tropes. In the biological world, species are developed by means of evolution, and on the basis of genetic mutations it is impossible to talk about identical characteristics of descendants with ancestors. It should be remembered, however, that all macro-objects are composed of basic units and theses are already qualitatively identical.

Let's take a look into the world of elementary particles based on the empirical findings of science. The fundamental electron particle, labelled as e^-, has the basic properties: charge -1, mass 0.5109989461(31) MeV, spin ½. If some other particle has some of the same properties as an electron but differs in its mass, the value of which is 105.6583745(24) MeV, then it is already another object, which in science is called a muon, denoted as μ^-.[3]

[3] It must be admitted, however, that the values in brackets (31) and (24) express measurement errors; the uncertainty has a value of ± expressed in the appropriate decimal place where the bracket is located. There must always be errors in measurement, but I consider these errors to be external effects working on objects that are not relevant in classifying them into kinds. For more details on measuring the mass of electrons see the more technically demanding article (Sturm and Köhler and Zatorski, et al. 2014); the authors from the Max Planck Institute for Nuclear Physics in Heidelberg used the Penning trap, a homogeneous magnetic field. The very precise

If we come across objects with the same charge, mass and spin properties, we always call each such object an electron. If it already had a different charge, or spin or mass, it can no longer be an electron, but something else. It is an essential property of electrons that they have the charge, mass, spin and other properties that are characteristic of all electrons.

Why do electrons exist in the world? Because there are basic universals that form them. If possible worlds existed and in them all the particles that we had become familiar with and which would have those same properties and only those as our electrons and would behave like electrons, then we could declare that electrons also exist in these worlds. In all possible worlds, these objects, under the given circumstances, *ceteris paribus*, behave the same.

When a scientist measures particle 1, which has charge of -1, a mass of 0.5109989461 MeV and spin of ½, and again manages to measure the occurrence of another particle 2, which has a charge of -1, a mass of 0.5109989461 MeV and a spin of ½, he declares that in both cases he measured properties indicative of an electron. What is the difference between the first measured particle e^{-}_{1} and the second measured particle e^{-}_{2}? A tropist would say that we have two very similar objects before us, which we call by the common name of electron. A realist would say that we have before us two qualitatively identical objects. We can decide which statement is more philosophically correct by comparing the properties of both numerically different objects e^{-}_{1} and e^{-}_{2}.

$$\boldsymbol{e^{-}_{1}}\ \mathbf{x}\ \boldsymbol{e^{-}_{2}}$$

-1 x -1

0.5109989461 MeV x 0.5109989461 MeV

½ x ½

We have expressed the properties of both measured particles numerically. A tropist would say that the property of 0.5109989461 MeV of particle e^{-}_{1} is only very similar to the 0.5109989461 MeV property of particle e^{-}_{2}. But

resulting value of the mass of the electron in units of the ion's mass (Sturm and Köhler and Zatorski, et al. 2014, 469) is m_e = 0.000548579909067(14)(9)(2). Charge, spin and mass should be regarded as intrinsic, non-relational properties of microscopic particles (Dorato 2006, 144).

I do not see any similarity here, but sameness! Therefore, in the world of elementary particles, realism establishing universals that science can also express numerically holds the upper hand.

5. Chora or the Physical Ether as the Bearer of Bundles

I think that bundle theories of universals have a fundamental problem explaining the *immanence* of universals themselves. If there were only universals without their instantiation, there could not be numerically multiple objects with the same properties–at most, there would be only one exemplar of each kind in a single Platonic heaven. There could be no “here” or “there.” For instance, Lafrance (2015) postulated a space in which universals are instantiated and their numerical distinctness is secured by spatio-temporal coordinates (cf. Hawthorne and Sider, 2002). I cannot identify with such projects, since they may lead to idealism or, at best, have only a phenomenological status. In the conception of dispositional universals, an energetic source is required to enable the manifestation of the universals’ own potentials; at the same time, it assigns them identity, differentiates them from one another, and grants them existence.

In order for one universal to act upon another, it requires a source of energy, which may be the all-pervading ether. A notion very close to what is needed here was considered by Plato in the *Timaeus*, namely the concept of *chora* (χώρα), understood as a receptacle capable of receiving forms. To explain what existed before the generation of the world, Plato posits a threefold distinction: “We may indeed use the metaphor of birth and compare the receptacle to the mother, the model to the father, and what they produce between them to their offspring...” (*Timaeus* 50d). The concept of *chora* is notoriously difficult to translate, as the word has no meaning; it is, as Sallis (1999, 115) puts it, “intrinsically untranslatable.” Based on interpretations of Plato’s text, and for the purposes of this text, we may accept that *chora* is “both space and matter all at once” (Zeyl 2010, 118); see also Jelinek (2015, 13, 22), Sallis (1999, 153), and Zeyl and Sattler (2023, chap. 6). The reasons for this spatio-material unity of the receptacle are two:

> First, we think that spatio-temporal particulars must be made up of *something.* They cannot be mere constellation of properties, mysteriously bundled together and even more mysteriously capable of maintaining that bundling as they move through space. And they cannot be made up of space, if space is sheer emptiness. (Zeyl 2010, 118)

Chora is formless and without properties, because it receives forms, just as manufactures of scents prepare the liquid to be as odourless as possible so the scent stands out, and those making impressions smooth the surface to clearly receive the imprint. (*Timaeus* 50e–51a). We shall not by wrong if we describe chora "as invisible and formless, all-embracing, possessed in a most puzzling way of intelligibility, yet very hard to grasp" (*Timaeus* 51a-b). Chora is filled with forces that exist in a state of disequilibrium; it is a kind of necessity, but a disordered one. "It is rather a necessity that would operate outside the law... resisting the rule of νοῦς even if responsive to its persuasion. This necessity is also called the *errant* form of cause" (Sallis 1999, 92).

The pre-elemental qualities of *chora*—water, fire, earth, and air—shake *chora* itself, and in turn, *chora* shakes them,

> with the result that they came to occupy different regions of space even before they were arranged into an ordered universe. Before that time they were all without proportion or measure; fire, water, earth and air bore some traces of their proper nature, but were in the disorganized state to be expected of anything which god has not touched, and his first step when he set about reducing them to order was to give them a definite pattern of shape and number (*Timaeus* 53a–b).

In this way, the Demiurge gave the world its order, but with the qualification that a certain degree of imbalance remains present within the cosmos. He brings about a state of near-uniformity, because susceptibility to agitation remains within the unalterable aspect of necessity (ἀνάγκη), which can be persuaded to serve the ends of intellect (νοῦς) (Zeyl 2010, 126).

Zeyl (2010, 122) conceives of chora as analogous to the sea, understanding it as both matter and space. We should imagine the pre-elements as waves in the ocean. Water is the material substratum, and the waves are spatial particulars. What differentiates one wave from another is its location in space and time, as well as its physical configuration (cf. Jelinek 2015, 23).

The entire model of chora proves to be highly useful for a bundle theory of universals. Universals may be instantiated in the underlying matter, which is chora. Universals are like waves: merging, colliding, or cancelling one another out. All particulars are waves of the sea. Particulars are thus bundles of universals that mutually interact thanks to a shared substratum of which all such bundles are a part. When the material of an object is exchanged or transferred into another, the identity of either object need not change, what is relevant is the form, the stability of the bundle.

Perhaps a more modern name for this matter, Plato's chora, could be found in the *ether* of the physicists. The concept of ether has undergone several centuries of development; in the twentieth century, it was even discredited by the experiments of Michelson and Morley, but it is today reemerging. The *ether* was proposed by Aristotle in his model of the supralunar world as a fifth element, considered to be pure, unchangeable, imperishable, and unfathomable. This model persisted until the seventeenth century, when a mechanistic image of the world replaced it. In the Cartesian worldview, the motion of bodies no longer depended on their intrinsic nature, but on a subtle matter that mediated the action of one body upon another (de Andrade, Faber and Rosa 2013, 560, 562).

Newton even speculated that his occult force of gravity is, in fact, a most subtle Spirit diffused throughout space (Newton 1846, 507), which would explain action at a distance. Later, Faraday employed the concept of ether in his theories; according to him, all space was filled with lines of force, and the atoms of matter were conceived as centers of force. Maxwell's ether was the indispensable medium for describing the phenomena of electromagnetic theory, including the wave propagation of light (de Andrade, Faber, and Rosa 2013, 560, 562).

The situation changed with the advent of special relativity, although with general relativity, substantialist tendencies re-emerged to restore a role for ether, now in the form of space-time acting on the distribution of matter. The revival of the concept of ether as *plenum* is evident in quantum mechanics, under notions such as quantum vacuum or dark energy. It has been recognised that “the concept of ether is making a strong comeback in physics” (de Andrade, Faber, and Rosa 2013, 573).

Speculation may arise as to how to define *ether* precisely. Is it an atomless *gunk*, that is, a substance infinitely divisible into ever smaller parts? (cf. Lewis 1991, 20; Sider 1993, 286). Neither logically nor empirically is such a possibility excluded; discrete quanta at the value of Planck’s constant may depend on form, while the matter composing them may be infinitely indivisible.

The exact nature of the matter in which the entire scenography of the world unfolds may be unknown to us. And yet, as in the case of chora or ether, it may hold that the substratum itself cannot be defined precisely because it is formless.

6. Dispositional Universals–Basics Premises

I postulate a basic axiom from which an entire proposed metaphysical model of the world will be based:

1) The world is made up of basic dispositional universals.

What are basic dispositional universals? From a philosophical point of view, I can only say that *some.* From a scientific point of view, these are basic properties, such as charge, mass and spin. Independent new particulars, such as fundamental particles quarks, leptons and bosons, are assembled from these universals. More complex elements are assembled from these, i.e., mesons and baryons are assembled from quarks. By combining a proton, a neutron and an electron hydrogen is created. By combining more protons, neutrons and electrons, the other elements in the periodic table are formed. All elements are already more complex universals with specific properties. How is it possible that by combining basic universals a new object is formed?

Basic universals are dispositional in character; therefore, I propose calling them dispositional universals. Universals themselves are potencies, and power gives them ether, as I claimed in section 5. So, the universal and ether together give rise to a dispositional universal. In dispositional universalism, the dispositional universal is manifested under suitable stimuli; if one hydrogen H meets another H (in the presence of a given energy, distance and angular motion), both have the ability to bind to one another; they are stimuli for each other, and a stable H2 molecule arises, with its stability ensured by the mutual manifestation of both Hs. This bundle holds together because it is bound by its very dispositional nature. This solves the problem of co-instantiation of bundle theory. An object is a bundle of basic dispositional universals, which by their manifestations already form a new dispositional universal of a higher order.

We can now formulate another thesis:

2) Basic dispositional universals are connected into bundles and thus create new objects through their mutual manifestations.

Complex elements can in turn connect into even more complex structures, such as when H2 combines with oxygen O and together they manifest a new stable bundle of H_2O. Physics textbooks describe how elementary particles hold such a bundle together with their energetic bonds. For example, Ingthorsson, in his powerful particulars view of causation, postulates *interaction* as a fundamental relational component between particulars. All interactions occur between the constituent parts of objects and serve as the glue that binds those parts together into a unity. There are four fundamental types of interaction–the same as those identified by physics: strong, weak, electromagnetic, and gravitational. Differences in the strength of these interactions allow us to distinguish between different physical systems, each displaying a different kind of unity depending on the interactions by which it is constituted (Ingthorsson 2021, 103–105).

We are, of course still moving in the area of science, and this may turn out to be wrong in time. Therefore, to state that these exactly are the basic universals and these are their combinations is only an empirical fact. From a philosophical point of view, the concept of dispositional universals is interesting for us, as it solves many problems of philosophy. How many kinds of basic dispositional universals exist cannot be answered exactly; we can only stick to the postulates of science and its incidental findings.

Some findings of science about the ontological nature of fundamental particles and their dispositional manifestations may be controversial or difficult to explain, for example, due to their indeterministic character, complementary dualism, quantum entanglement. The biggest stumbling block of science is the inability to connect the micro world (the quantum world) with the macro world (Newton-Einstein world).

3) Dispositional universals of a higher order are manifested in different behaviours than the dispositional universals of a lower order from which they are composed.

The bundles from which objects are assembled are mereological sums of parts and they are in fact an entirely new wholes. If 3) were false, then it would commonly happen that we could pass with our bodies through walls from one room to another. An elementary particle is able to tunnel through a barrier, but the entire object composed of these particles cannot, or it is highly unlikely.

The whole is thus more than the sum of its parts. A bundle of dispositional universals forms a new substance whose parts are dependent on the whole and the whole on the parts. A liver cannot exist without an organism, just as the organism cannot exist without the liver. The whole is held together by its necessary dispositional universals, which may be joined by accidental dispositional universals. A universal is necessary only for the given context of the whole. The liver is a necessary component of the whole; tooth decay is its accidental addition. The border between necessary and accidental universals is not always clear. Eye colour can be an accidental property, but it is causally determined by the DNA chain. At the same time, it applies that a person would be a person even without a specific eye colour, but without their fully functioning development we are already thinking about a defect and not a fully developed human nature.

Wholes often manifest differently than their parts. When we move a chair, we are moving the entire inertial frame of reference in which the electrons continue to rotate around their nuclei. The level of interconnection of parts in the whole can be graded, from simpler connections to stronger bonds. The piling up of snow on a roof can be the first example. One flake joins another, and together they form an increasingly heavy mass, until finally the roof collapses. The second example can be a person, whose parts create an interconnected organic unity.

Transitions from the microworld, through molecules, objects, living creatures, communities, planets, galaxies and universe/s, lead to various leaps.

At the lowest levels, the world can run continuously[4] but between different levels we can see jumps. A new snowflake that falls on a roof will cause a jump from the standard appearance of a persistent roof to its collapse. The emergence of new qualities can be compared analogously to the weather; they are governed by synergistic processes, attractors; air flow jumps to a new level into a new object, a tornado. Likewise, a husband in a household shows standard care expressions, but he comes to the crowd and is overcome by crowd psychosis and now shows aggressive manifestations determined by the whole–the crowd.

4) Anything cannot be connected to anything.

Wholes cannot be assembled from arbitrary universals–this is the problem that troubles the bundle theory of powerless universals. It is not possible, for example, for a triangular universal to connect with a quadrilateral universal, just as it is equally not possible, *ceteris paribus*, for the two negative poles of a magnet to be non-forcefully connected, or for a person with damaged speech organs, *ceteris paribus*, to have the ability to recite *The Iliad* aloud. In the ontology of dispositional universals connections between universals are constrained by their potencies and mutual stimulus conditions.

5) Anything cannot be manifested in any way (What is logically conceivable is not always metaphysically possible).

It is not metaphysically possible for a donkey's DNA to manifest itself through human speech. Or that a possible world exists in which Bertrand Russell is a stupid jellyfish, or salts do not normally dissolve in water. Dispositional universals are manifested under suitable stimuli, always necessarily the same, *ceteris paribus*, in all possible worlds. Why this is so is an empirical problem. At least on an inductive level, we find out how given kinds of objects are manifested under given circumstances, but *why* it is in their nature to manifest in this way is a more demanding question.

6) Objects appear to us through their manifestations of dispositional universals, and we define them on the basis of these manifestations.

[4] According to science, in reality discretely delimited by Planck's constant.

Objects appear to us through their dispositional manifestations. A blue surface reflects light with a wavelength in the range of 450 nm to 495 nm. The same identity relationship holds between blueness and wavelength as between water and H_2O.[5] When someone perceives blue at a wavelength of 450 nm and someone else perceives a colour at a wavelength of 450 nm, we can say that they are both perceiving the same manifestation of blueness. A wave that spreads from the surface of the object is this universal, and the surface of the object is dispositionally arranged to reflect light of this wavelength. Two surfaces that emit an identical wavelength exemplify the same universal, blueness – which is the ability of elementary particles occurring on the surface to absorb and then emit photons of the given wavelength. The wave is available to be manifested as a sensation upon stimulation with the perceiver's eyes and mind.

We identify objects on the basis of their manifestations; we classify them into kinds and create concepts and definitions about them. At the same time, the necessary properties of objects should be captured in the definitions. Gold, an object with a proton number of 79, always manifests the same under given circumstances (Kripke 1980), and its specificity is, for example, malleability and solubility in aqua regia. It is chemically stable and therefore does not oxidize or corrode under normal conditions. And so on. The semantics of words is created based on dispositional manifestations of the kinds of objects that are represented by symbols in our vocabulary.

An object isolated from any interactions with the world does not make itself known; the world is therefore an interconnected set of interactions that trigger each other's dispositions. "The world is a single whole, composed of properties whose essence and identity are determined by their place in that whole" (Mumford 2004, 184). A property is to be nothing more than a set of connections to, and causal powers for, other properties (Mumford 2004, 185).

If, for example, gold occurred in another universe where there were alien universals, then when in contact with these alien universals, it could show alien manifestations, and we would devise other definitions about it. Even in our world, everything does not have to interact with everything; therefore, we may not have recognised all the manifestations of known kinds, or

[5] On the relation of identity, see (Kripke 1971).

these manifestations have not yet been identified for us. More than once, science has come up with new kinds of manifestations; for example, the discovery of electric power, the power obtained by splitting the atom, the collapse of massive objects into the form of black holes, and others.

7) Dispositional universals are modal in nature.

Point 7) is limited by point 5). But this is the greatest mysticism of dispositional universals and dispositional theories in general. I rejected the laws of nature as ideal entities, "free floaters," because according to dispositional theories, the world is governed by the very nature of the properties of matter. This is why the regularities must be sought in dispositional universals, from the most basic up to organised new units. An object in itself contains the possibilities[6] of how it would necessary manifest itself under the given circumstances and these possibilities are within it, and we can think of them as records on a strip of film. The stimulus illuminates the given image and it then manifests in the form of force action. Sometimes the manifestation means the expiry of the object, like the manifestation of the breakability of a vase–the mutual energy bonds of the object's properties are released and it disintegrates–and sometimes the manifestation persists permanently, as the gravitational force of a star.

Where these modalities are located is a debatable question. We can sense brownness, firmness or the curved silhouette of a vase, but not its breakability. In this case, sensory experience is insufficient; we have contact only with the manifestations of objects and not with the capabilities that are in the potential of their realisation. For this reason, when examining the world, we must help ourselves with rational experience; this means, we infer the existence of potencies in particulars belonging to the same kinds from empirical experience with the manifestations of given kinds of objects. Belief in the validity of our predictions is possible because the given objects fall into the given kinds, and for each kind, such and such manifestations are characteristic under the given stimuli. The heating of a wax form can take on almost infinite possibilities of its geometric arrangement, and all these

[6] B. Vetter prefers the term potentiality, they are "possibilities rooted in objects; they are like possibilities, but they are properties of individual objects" (Vetter 2015, 3).

possibilities must be contained in the entire wax. If no suitable stimuli are present for manifestation, then these manifestations slumber in objects as unrealised possibilities. Their manifestations depend on the possible arrangements of each property in relation to others, and these possible arrangements must be included in a disposition as its necessary reactions to those possible stimuli. Empirical manifestations of objects are dependent on the evolution of the entire universe; therefore, the universe is also reflected in them, as in Leibniz's monads, and not only the universe, but also all metaphysically possible configurations of metaphysically possible worlds. The movement of a comet is dependent on the overall structure of the universe and its past; this comet could show different manifestations, if it were found in different conditions–in different worlds.

Potential manifestations of dispositions are not merely something exclusively in us that we attribute to objects; they are natural properties in nature. Therefore, there is a difference here between the logical and metaphysical possibility. But what is logically possible does not have to be metaphysical. The world is ruled by *de re* necessity. All the possibilities of how the world can be metaphysically are reflected in objects, because each object manifests itself in a specific way with every other object, and these manifestations are within its potential. The possibilities of development of the world are limited by the possibilities of manifestations from mutual interactions of the dispositional universals it contains. If parallel universes existed and were to intermingle, we could recognise new manifestations of familiar objects. If everything comes from one Multiverse and is of one essence, then there should be here the possibility for mutual interaction of all objects within their permissible metaphysical possibilities. These unrealised possibilities should be real existing forces that are part of dispositional universals. For us, they are almost mystical entities, because their number of manifestations can reach enormous dimensions.

Where these unrealised possibilities are located is still a great mystery of philosophers of dispositional properties. How can something exist when it is not current? Where is the explosiveness of the grenade located when the grenade has not yet exploded? Answering this question is very difficult, but in this case, universalists have an advantage over tropists. Platonic realism offers a robust solution to the problem of unmanifested dispositions by grounding dispositional directedness in a relation between abstract universals that exist independently of any instantiation. “There can be such a relation because the manifestation universal can still exist even if the particular’s disposition never actually manifests” (Tugby 2013, 461). In contrast, Aristotelian immanent realism requires that the manifestation universal be instantiated somewhere in space and time in order to secure directedness. Thus, while Platonism appeals to the pure existence of universals, Aristotelianism ties them to actual or possible occurrences within the concrete world (Tugby 2013, 461).[7]

In quantum physics, the wave function ψ is the carrier of mutually exclusive states: the decay or non-decay of uranium. According to Everett’s (1957) concept of quantum physics, the many worlds interpretation, different realisations of ψ split into parallel branches. For the eigenfunction of observation $\phi^{S}{}_{1}$ corresponds to the state where uranium has not decayed, and $\phi^{S}{}_{2}$ corresponds to the state where uranium has decayed. However, from the viewpoint of the Multiverse, this process still represents one universal of uranium, in which all branching states are included. Ultimately, this is not about ontological potencies in the traditional sense, because all possible outcomes of the wave function are realised in separate branches of the Multiverse.

Everett’s concept of split worlds also brings several philosophical problems:

> Does the pointer itself split in two? Or are there two numerically distinct pointers? If the whole universe splits into two, doesn’t this wildly violate conservation laws? There is now twice as much energy and momentum in the universe than there was just before the measurement. How plausible is it to say that the entire universe splits? (Ney 2013, 33).

Therefore, we do not have to think about split worlds, but about the *parallel* course of physical worlds in which all metaphysically possible states are realised. These worlds may be patterns in the one universal quantum state that emerge as the result of its evolution; distinct components of the quantum state come to evolve independently of one another (Ney 2013, 34).

[7] Tugby (2013) argues only in favour of Platonism. He could explain his theory by refusing to answer “the much-discussed question of how the understand the relationship between universals and their concrete instantiations” (Tugby 2013, 452).

Appealing to possible worlds is only an aid in the era of still prevailing logical positivism, in which we need truth-makers even for propositions about unrealised possibilities of dispositions. I am rather an advocate of the assumption that dispositional universals are an unexplored dynamic mechanism full of potentialities, whose empirical immediate dimension escapes us. We know from inductive experience that possibilities exist in some form. From the manifestations of the given kinds of objects ("These metals expanded with heat") we rationally ascribe possible manifestations to the same kinds ("This is a metal and will also expand with heat, because all metals have the ability to expand when heat is applied").

Dispositional universals may also be "metaphysical algorithms" whose outputs depend on the inputs. Similar to an electronic calculator that does not need to contain all combinations of states, it only needs the function (_ + _ = _) into which values are filled in, nor do dispositional universals need to contain the information that S. Mumford is possibly married to N. Cartwright. It is sufficient that S. Mumford's metaphysical algorithm allows him to do so. Just as a calculator is able to calculate 18325 + 12365, so, too, might S. Mumford be able to marry N. Cartwright, if the right (and even advantageous) circumstances arise. The calculator does not endlessly count somewhere in the range of 18325 + 12365, and S. Mumford does not keep saying "Yes!" to N. Cartwright. Dispositional universals are forms that shape the world according to current inputs of another dispositional universals on the basis of what they are capable of by their construction.

8) Kinds exist at all levels of arrangement of dispositional universals.

From the basic kinds of universals, larger units, such as particles, atoms, molecules, proteins, plants, viruses, bacteria, insects, animals, people, society... are assembled. We would find common universals at every level. Common universals that characterise people are, for example, the command of language–grammar, logical concepts (and, or, not, greater/lesser, part/whole...), expression of emotions (smiles, frowns), music, dances, child-care of mothers, incest avoidance (Brown 2004, 48–51).[8] But the more we move to more complex levels of arrangements, the number of kinds grows and our

[8] In my article (Károly 2024), I proposed a method for detecting features of human nature based on discomfort.

classification of particulars into kinds also becomes more complicated – complications are caused not only by the cladistic tree in the animal kingdom, but also by determining which properties are considered essential and why specifically these. We cannot deal with the problems faced by biologists in this text. I can only summarise the whole problem by saying that all objects are made up of kinds of lower levels, which in a bundle create a new particular and a new kind. Common universals of lower levels characterise kinds at a higher level; for example, the dispositional universal the ability to learn human speech occurs in all healthy individuals, the whole of which we call humanity.

9) If world events are conditioned by dispositional universals, then laws of nature do not exist.

The implication in 9) would be true even if the antecedent were false. The traditional humeans reached the same conclusion (world events are not conditioned by anything), as did the tropists (world events are conditioned by dispositional tropes). Non-humean philosophers who claim that laws of nature do not exist are, for example, Martin (2008), Mumford (1998; 2004), Bird (2007) and Ellis (2001). The tropist Martin claims that laws of nature appear to be ontologically otiose, because “If you accept arguments for a realism of dispositions and their reciprocal disposition partners and grand the dispositions could be fully actual although their partnerings or manifestings might not be, then *what* is the need for universal law?” (Martin 2008, 22). According to Mumford, an advocate of universals: “If necessity resides in the propertied particulars in nature, there will be no need for laws. Particulars are powerful in virtue of their properties. They are not powerless discrete units so do not require laws to make them act. Immanent necessity in nature might then become a good reason why there are no laws of nature” (Mumford 2004, 63).

Point 1), from which I gradually formulated theses 2) – 8), can be reformulated into a statement that is the main statement of dispositional universalism:

10) World events are conditioned by dispositional universals.

Then from 9) and 10) we get the final conclusion:

11) Laws of nature do not exist.

Of course, this is not a new conclusion; it is a universal claim among dispositionalists. What is new is my proposed ontology of dispositional universals and their basic behaviour, as outlined in the ten premises.

7. Statements about the Lawfulness of Dispositional Universals in the D-N Model

What effect do statements 10) and 11) have on the D-N model we mentioned in the introduction? Hempel says that dispositional explanations do conform to the covering-law conception of explanation. Disposition is a property dealt with by a theory, and explanatory principle is part of what the theory asserts about disposition and, as general theoretical principle, it expresses a nomic claim (Hempel 1974, 374).

Statements in dispositional universalism about manifestations of dispositions have a nomic character. The statement about a particular[9] "If someone strikes this breakable glass vase, *ceteris paribus*[10], then this vase will break" is valid because this particular is represented by the kind to which all the breakable glass vases that were, are, will be and could be belong. If a specific kind of breakable vase is mass-produced in a production line, it is enough to break just one with a hammer, and we can conclude for all these products: "For every breakable glass vase of this kind, if someone hits it with a hammer, *ceteris paribus*, then the vase will break." If we would like to explain why the vase broke in the factory through the D-N model, we would state:

C_1: This glass vase is breakable (It belongs to the kind of breakable glass vases).

C_2: After this vase was struck with a hammer, *ceteris paribus.*

L: For each breakable glass vase of this kind if someone hits it with a hammer, *ceteris paribus*, the vase will break.

E: The vase broke.

[9] In our context, it is a bundle of dispositional universals.

[10] Nobody prevents a vase from breaking by strengthening the structure of the glass, for example by freezing it.

In this case *L* is our inductive generalisation; it does not capture the law as something ontological. Rather, in the case of *L* the name lawfulness, which is closer to the meaning of the universal manifestation of objects, would be more appropriate. The lawfulness of dispositional manifestations can also be expressed more generally, for example: "For all breakable bodies D, it is true that if they are acted upon by force F, they will break."

If dispositional universalism is true, then the problem of induction occurring because of the empirical failure of our knowledge. Likewise, various deviations in measurement are the results of various environmental influences, which can be difficult to avoid; therefore, the effort of science is to define the given kind in the presence of the smallest possible external disturbing elements. If dispositionalism is true and universals exist, then we do not need to perform endless experiments to establish the truth of the proposition "If *F*, then *G*" nor succumb to Humean scepticism, because, as Demarest claims (2017, 48):

> Scientists need only perform a relatively small number of experiments on a single kind of particle before they feel confident that they have captured its true nature, or, in my terms, the essential dispositions of its potencies.

If some object occurs with an atomic number of 79, then it must always manifest itself as gold; this is metaphysically inevitable. Therefore, even nomic statements about the manifestations of gold must be universally true. A law is already an additional generalisation that we humans create from knowledge of dispositional universals in the presence of given stimuli and *ceteris paribus*. Laws are expressions of the lawfulness of dispositions.

8. Conclusion

To repeat, we will again recall the eleven points that characterise the nature of dispositional universals and thus the nature of the functioning of the world.

1) The world is made up of basic dispositional universals.
2) Basic dispositional universals are connected into bundles and thus create new objects through their mutual manifestations.

3) Dispositional universals of a higher order are manifested in different behaviours than the dispositional universals of a lower order from which they are composed.
4) Anything cannot be connected to anything.
5) Anything cannot be manifested in any way. (What is logically conceivable is not always metaphysically possible.)
6) Objects appear to us through their manifestations of dispositional universals, and we define them on the basis of these manifestations.
7) Dispositional universals are modal in nature.
8) Kinds exist at all levels of arrangement of dispositional universals.
9) If world events are conditioned by dispositional universals, then laws of nature do not exist.
10) World events are conditioned by dispositional universals.
11) Laws of nature do not exist.

Dispositional theorists arrive at the conclusion that laws of nature do not exist. This means that the world is not governed by some ideal laws from above, but the lawfulness itself comes from the dispositional nature of objects. According to the theory I propose, objects are bundles of dispositional universals instantiated in physical ether, like the traces of seal in wax. Dispositional universals govern the world by virtue of their metaphysical nature.

Acknowledgements

I thank prof. Matúš Porubjak for guiding me, in personal discussion, toward the idea of introducing *chora* as the bearer of bundles of universals, as an alternative concept to the physical *ether*.

Funding

The article was written under the support of FPPV-03-2023.

References

Armstrong, David. M. 1978. *A Theory of Universals*, volume 2. Cambridge: Cambridge University Press.

Armstrong, David. M. 1985. *What is a Law of Nature?* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316499030

Armstrong, David. M. 1989. *A Combinatorial Theory of Possibility.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139172226

Bird, Alexander. 1998. "Dispositions and Antidotes." *The Philosophical Quarterly,* 48 (191): 227–34. https://doi.org/10.1111/1467-9213.00098

Bird, Alexander. 2007. *Nature's Metaphysics. Laws and Properties.* Oxford: Oxford University Press, 2007. https://doi.org/10.1093/acprof:oso/9780199227013.001.0001

Bromberger, Sylvain. 1966. Why Questions. In *Mind and Cosmos*, edited by Robert R. Colodney, 86–111. Pittsburgh: University of Pittsburgh Press.

Brown, Donald E. 2004. "Human Universals, Human Nature & Human Culture." *Daedalus*, 133(4): 47–54. http://www.jstor.org/stable/20027944

Choi, Sungho, and Michael Fara. 2021. "Dispositions," *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), edited by Edward N. Zalta. Last updated Jun 22, 2021. https://plato.stanford.edu/archives/spr2021/entries/dispositions/

de Andrade, Elaine Maria Paiva, Faber, Jean and Luiz Pinguelli Rosa. 2013. "A Spontaneous Physics Philosophy on the Concept of Ether Throughout the History of Science: Birth, Death and Revival." *Foundations of Science*, 18: 559–77. https://doi.org/10.1007/s10699-013-9336-9

Demarest, Heather. 2017. Powerful Properties, Powerless Laws. In *Causal Powers*, edited by Jonathan D. Jacobs, 38–53. New York: Oxford University Press. https://doi.org/10.1093/oso/9780198796572.003.0004

Descartes, René. 1996. *Meditations on First Philosophy.* Cambridge: Cambridge University Press.

Dorato, Mauro. 2006. "Properties and Dispositions: Some Metaphysical Remarks on Quantum Ontology." *AIP Conference Proceedings.* 27 June 2006; 844(1): 139–57. https://doi.org/10.1063/1.2219359

Douglas, Heather E. 2009. "Reintroducing Prediction to Explanation." *Philosophy of Science*, 76(4): 444–63. https://doi.org/10.1086/648111

Dretske, Fred I. 1977. "Laws of Nature." *Philosophy of Science*, 44 (2): 248–68. https://www.jstor.org/stable/187350

Ehring, Douglas. 1997. *Causation and Persistence. A Theory of Causation.* New York: Oxford University Press.

Ellis, Brian. 2001. *Scientific Essentialism.* Cambridge: Cambridge University Press.

Everett III, Hugh. 1957. "'Relative State' Formulation of Quantum Mechanics." *Reviews of Modern Physics*, 29(3): 454. https://doi.org/10.1103/RevModPhys.29.454

Hawthorne, John, and Theodore Sider. 2002. "Locations." *Philosophical Topics*, 30(1): 53–76. http://www.jstor.org/stable/43154380

Heil, John. 2003. *From an Ontological Point of View.* Oxford: Oxford University Press. https://doi.org/10.1093/0199259747.001.0001

Hempel, Carl G., and Paul Oppenheim. 1948. "Studies in the Logic of Explanation." *Philosophy of Science*, 15(2): 135–75. http://www.jstor.org/stable/185169

Hempel, Carl G. 1974. "Dispositional Explanation and the Covering-Law Model: Response to Laird Addis." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1974: 369–76. http://www.jstor.org/stable/495813

Hume, David. 1960. *A Treatise of Human Nature.* Oxford: Clarendon Press.

Hildebrand, Tyler. 2023. *Laws of Nature.* Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009109949

Inghtorsson, R. D. 2021. *A Powerful Particulars View of Causation.* New York and London: Routledge.

Jansen, Ludger. 2009. Aristotle's Theory of Dispositions: From the Principle of Movement to the Unmoved Mover. In *Debating Dispositions: Issues in Metaphysics, Epistemology and Philosophy of Mind*, edited by Gregor Damschen, Robert Schnepf, and Kasrten Stüber, 24–46. Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110211825

Jelinek, Elizabeth. 2015. "An Examination of Plato's Chora." *Environment, Space, Place*, 7(1): 7–27. https://doi.org/10.5840/esplace2015711

Károly, Tomáš. 2024. "The Universals of Human Nature: A Method of Their Detection through Scientific Evidence and Literary Fiction." *Pro-Fil*, 25(2): 51–65. https://doi.org/10.5817/pf24-2-39121

Kripke, Saul A. 1971. Identity and Necessity. In *Identity and Individuation*, edited by Milton K. Munitz, 135–64. New York: New York University Press. https://doi.org/10.1017/S0012217300030122

Kripke, Saul. A. 1980. *Naming and Necessity.* Cambridge, Massachusetts: Harvard University Press.

Lafrance, Jean D. 2015. "A Bundle of Universals Theory of Material Objects." *The Philosophical Quarterly*, 65(259): 202–19. https://doi.org/10.1093/pq/pqu078

Leibniz, Gottfried W. 1985. *Theodicy: Essays on the Goodness of God, the Freedom of Man, and the Origin of Evil*, edited with an introduction by Austin Farrer. La Salle, Illinois: Open Court Publishing Company.

Leibniz, Gottfried W. 1989. The Monadology. In *Philosophical Papers and Letters*, edited by Leroy E. Loemker, 643–53. Dordrecht: Kluwer Academic Publishers.

Lewis, David. 1986a. *Philosophical Papers.* Volume II. Oxford: Oxford University Press.

Lewis, David. 1986b. *On the Plurality of Worlds.* Oxford: Blackwell.

Lewis, David. 1991. *Parts of Classes.* Oxford: Blackwell.

Lewis, David. 2001. *Counterfactuals.* Oxford: Blackwell.

Loux, Michael J. 2006. *Metaphysics: A Contemporary Introduction*, 3rd ed. New York: Routledge.

Malebranche, Nicolas. 1997. *The Search after Truth: With Elucidations of The Search after Truth.* Cambridge: Cambridge University Press.

Martin, C. B. 1994. “Dispositions and Conditionals.” *The Philosophical Quarterly*, 44(174): 1–8. https://doi.org/10.2307/2220143

Martin, C. B. 2008. *The Mind in Nature.* New York: Oxford University Press.

Molnar, George. 2006. *Powers. A Study in Metaphysics*, edited by Stephen Mumford. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199204175.001.0001

Mumford, Stephen. 1998. *Dispositions.* Oxford: Oxford University Press.

Mumford, Stephen. 2004. *Laws in Nature.* New York: Routledge.

Newton, Isaac. 1846. *The Mathematical Principles of Natural Philosophy.* New York: Daniel Adee.

Ney, Alyssa. 2013. Introduction. In *The Wave Function: Essays on the Metaphysics of Quantum Mechanics*, edited by Alyssa Ney and David Z. Albert, 1–51. Oxford: Oxford University Press.

Ott, Walter. 2009. *Causation and Laws of Nature in Early Modern Philosophy.* Oxford: Oxford University Press.

Planck, Max. 1963. *The Philosophy of Physics.* New York: Norton Library.

Plato. 1977. *Timaeus and Critias*, translated with an introduction by Desmond Lee. London: Penguin Books.

Putnam, Hilary. 1973. “Meaning and Reference.” *The Journal of Philosophy*, 70(19): 699–711.

Rosenberg, Alex, and Daniel W. McShea. 2008. *Philosophy of Biology: A Contemporary Introduction.* New York: Routledge.

Sallis, John 1999. *Chorology: On Beginning in Plato’s Timaeus.* Bloomington, Indiana: Indiana University Press.

Sider, Theodore. 1993. “Van Inwagen and the Possibility of Gunk.” *Analysis*, 53(4): 285–89. https://doi.org/10.2307/3328252

Sturm, S., F. Köhler, and J. Zatorski et al. 2014. “High-Precision Measurement of the Atomic Mass of the Electron.” *Nature*, 506: 467–70. https://doi.org/10.1038/nature13026

Tooley, Michael. 1977. “The Nature of Laws.” *Canadian Journal of Philosophy*, 7(4): 667–98. https://doi.org/10.1080/00455091.1977.10716190

Tugby, Matthew. 2013. “Platonic Dispositionalism.” *Mind*, 122(486): 451–80. https://doi.org/10.1093/mind/fzt071

Vetter, Barbara. 2015. *Potentiality: From Dispositions to Modality.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198714316.001.0001

Zeyl, Donald. 2010. Visualizing Platonic Space. In *One Book, The Whole Universe: Plato's Timaeus Today*, edited by Richard D. Mohr and Barbara M. Sattler, 117–30. Las Vegas: Parmenides Publishing.

Zeyl, Donald, and Barbara Sattler, "Plato's Timaeus", *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), edited by Edward N. Zalta, and Uri Nodelman, Last updated May 13, 2022. https://plato.stanford.edu/archives/fall2023/entries/plato-timaeus/

Organon F 33 (1) 2026: 33–74 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2026.33102

RESEARCH ARTICLE

The Reductive Roots of Mechanistic Explanations

Roque Molina Marchese*

Received: 28 August 2024 / Accepted: 10 August 2025

Abstract: Mechanistic explanations play an important role in the scientific understanding of the mind. Mechanistic explanations reflect the commitments of *mechanistic philosophy*. This philosophy stresses a methodology based on decomposition into constituent parts and the synergistic integration of their activities. The methodology provides the *identity types* required for the formulation of explanations targeting the explanandum phenomenon. The basic relationship between mechanisms and their target explananda is mereological. A system is in intrinsic *compositional* relation to the participation of its constituent parts in the system's overall functionality. The compositional relation is primarily *inter-level*. The stability of the system requires also the existence of *intra-level causation* within the boundaries of minor sub-mechanisms. In combining the compositional relation with intra-level causation mechanistic explanations would track *nonreductive* relations. Intra-level causation and inter-level compositional effects are also reflective of a mereological relation. The mind-brain system is an example of a mereological relationship and so according to this logic, the mind would be rendered irreducible. Combining both intra-level *causation* and inter-level material *constitution* is nevertheless *not* sufficient to warrant mechanistic explanations being non-reductive. By emphasizing identity types as the bearers of

* Charles University
https://orcid.org/0009-0006-7304-492X
Department of Philosophy and Religious Studies, Náměstí Jana Palacha 1/2, 11638, Prague 1, Czech Republic
roque.marchese@gmail.com

the mereological relation, mechanists in general regard special-sciences and their kinds as actually reducible to the subsuming mechanisms. Applied to the mind-body relation this means the reduction of mind to brain states.

Keywords: Abstraction; intra-level causation; explanation; mechanisms; reduction.

1. Introduction

This article explores the way mechanistic explanations and their application specially in the neurosciences relates to the character of the mind as informed by special sciences like psychology. I argue here that despite mechanistic philosophy offering a reading of the part-whole relationship between mind and brain as implying irreducibility for the mind it nevertheless fails because of its commitments to type identities. Many special sciences refer to the existence of kinds like subjective mental states, economic transactions and biological processes as *irreducible* to basic physical kinds and their properties (Beckage et al. 2013; Hermida and Ladyman 2025; Weiskopf 2011). Thus, mental states as those present in for example economic transactions would be irreducible to neuroscientific states and their properties. At the same time, these examples of irreducible kinds are grounded on the operations of individual brains. It is here where mechanistic explanations approach psychological and mental phenomena by way of referring to the constitutive character of neural mechanisms and their characteristic activities as responsible for the synchronous "here and now" realization of mental phenomena. A basic assumption is that mechanistic explanations track composing structures and their properties representing subvenient bases on which higher-order phenomena are *physically* realized. Further, mechanistic explanations are explanatory to the extent that they successfully identify these structures as cases of natural kinds and their properties representing *types*. Thus, the explanatory power of mechanistic explanations rests on the recognition of *type identities* as the grounds on which mechanisms function and work for their capacity to realize higher-order phenomena. In that sense, mechanistic philosophy is in spirit reductionistic. Still, depending on how we understand the defining characteristics of mechanisms there is a possibility to read mechanistic philosophy as implying non-reductivism for mental properties. Mechanistic philosophy in general understands the existence of a certain explanandum as the result of objects, their parts and activities organized in such a manner that in conjunction with their here-and-now *physical occurrence* realize the target explanandum (Machamer, Darden and Craver 2000). Non-reductivism adds to this picture the idea that despite the explanandum being ontologically based on the realizing mechanisms, because of properties and characteristics unique to its macro-based behavior, it remains nonetheless non-identical to those mechanisms. The combination of mechanistic philosophy and non-identities would then in principle make a case for mechanistic explanations being non-reductionistic. To work, this combination requires a view of mechanisms as part of a mereological relation characterized by both inter-level and intra-level properties. Usually, the neurosciences approach the mind mechanistically by assuming causation between different *levels* of organization, i.e., both mind and brain "located" at their own level in the hierarchy of natural phenomena. Moreover, being there a causal pathway between brain and mind we can develop an explanation of the mind in terms of physical realization by the occurrence of brain states. Nevertheless, many mechanistic models are also examples of a *mereological relation*, i.e., the reciprocal structural as well as functional integration between the system and its component parts. The mereological relation holds not only between the explanandum phenomenon and its subvenient mechanisms but also between *major mechanisms* and their composing *sub-mechanisms* at the level of the realization bases. A major mechanism then is an example of a constitutive and functional part of the system composed by more basic underlying structures and their functions, i.e., a sub-mechanism. Given the possibility causal pawers inherent to the system are restricted to the instantiation of functional roles at the *intra-level* of local sub-mechanisms we speak of a *horizontal explanatory relation* (Kaiser and Krickel 2017). This relation confronts with a vertically oriented *constitutive* relation between explanandum and mechanism. If correct, mechanistic explanations would have the virtue of being explanatory at the intersection of this horizontal and vertical confluence. Thus, they explain the phenomenon by means of an epistemology grounded on intra-level causal effects working *within* and along the sub-mechanisms, i.e., their horizontality, together with an ontology working "between" the sum effect of the constitutive major mechanisms i.e., verticality. "Horizontal" here then refers to causal pathways operating internally at the local constitution level of sub-mechanisms. Further, mechanisms explain their target phenomenon also by means of an ontological constitutive relation operating inter-level *between* mechanisms, i.e., their verticality. Additionally, horizontal causation operating at the sub-mechanistic level is also *conditioned* by the mereological relation of parts to their wholes and vice versa: i.e., the whole system has a modolutary effect on the causal profile of the sub-mechanisms. Furthermore, combining intra-level causation with material inter-level constitution renders the entire mereological system irreducible; the structures and properties occupying intra-levels and inter-levels are mutually integrated as parts and wholes. This is why psychological properties as systemic properties are not reducible to the elementary units of the entire system. Psychological properties are expressions of a constitutive mereology conditioned in part by internal causal effects proper to subsuming micro mechanisms and in part by the global effects of systemic macro properties. Moreover, local intra-level causal effects are not necessarily inherited by the global systemic behavior. The entire system as such is in itself *causally distinctive*; there are separate systemic causes not necessarily reducible to the horizontal local causes found at the level of sub-mechanisms (Romero 2015). It is then paramount to understand why causation in a system is mainly confined to the mechanistic "intra" level of organization. Any system operates through the composing activities of its parts. At the same time the operation of the parts is conditioned by the qualities and behavior of the system as such. Causal effects from the parts resonates at the systemic level and in addition globally based systemic effects resonates at the elemental level of constituent parts. Causation here is reciprocal. Think of Sperry's wheel (Sperry 1991, 1980). Here the wheel is in motion and so even its components. The material components of the wheel participate in grounding the mere possibility of being there a wheel in motion at all but the wheel as a global system in motion determines also the location and displacement of its components atoms in space-time. Causation as a systemic property is integrated into the whole system. At the same time it can and must be instantiated simultaneously by different local elements at different times, i.e., its synchronous character (Cabral et al. 2022). The local elements here correspond in the mereology of mechanistic systems to the existence of sub-mechanisms. Sub-mechanisms are then causally synchronous. In other words, they “happen” at once and they “contain” the causal efficacy of “local spots” in the system i.e., its components. It is the synchronous existence of the sub-mechanisms’ causal activity that together support the integrity of the system in its here and now condition. As a result, these sub-mechanisms operate on their own basis following their own internal “causal clock” and can be operationally recruited when so necessary for the instantiation a certain property in the system. At the same time, the system is characterized by its inherent robustness, resilience and graceful degradation of functionality. There is an inbuilt redundancy to the operation of sub-mechanisms that causally speaking is available to the needs of the entire system when so needed (Bernard 2023). The role of the subsystems then is to perform and realize causal roles that can be recruited for safeguarding the resilience and redundancy of the whole system. Thus, causation as a structuring and organizing power is not evenly distributed across the entire system but must be *localized* to the function of certain systemic units, i.e., sub-mechanisms. Moreover, given the integrity, redundancy and robustness of the system it means that individual sub-mechanical units are to a certain degree in principle dispensable. The system can still operate in the relative presence or absence of sub-mechanisms as long as the functional integrity of the entire system is secured. This is the primary reason why the instantiation of causal powers must be *multiply distributed* across different sub-mechanisms to support the integrity of the system. Moreover, seen from the perspective of neural structures every such sub-mechanism must also be *plastic* or adaptable enough to instantiate different causal profiles. Transferred to the way the brain realizes mental properties, this reflects the structural and functional malleability typical for neuronal network organization. This plasticity-based brain property represents how neuronal networks perform and carry out different computational operations required for the implementation of cognitive functions (Bassett and Sporns 2017; Bassett, Zurn, and Gold 2018). On the other hand, this intra-level characterization of the brain’s systemic organization is not the way mechanistically oriented neuroscientists interpret the explanatory relationship holding between cognitive function and brain physiology. Rather, they assume that mechanisms explain this relationship in virtue of instantiating *inter-level* causation responsible for the implementation of a specific cognitive function. This means that according to many mechanists the correct view of realization between a higher-order property and it subvenient bases is dependent on its "verticality". In other words, a phenomenon is cosntituted by the synchronous presence of its fundamental components. Yet, even if the idea of intra-level causal operations. In line with the scale free and multiscale character of brain function and organization gives some support for the irreducibility of cognition to brain physiology, this is not a sufficient reason to believe that mechanistic explanations in general can be non-reductive. The main reason is that an essential aspect for mechanistic models to make sense is that they both causally (horizontally) as well as constitutive (vertically), must rely on the existence of *Types* when formulating explanations. It is the very type as such as a basic and recurrent unit of oranization that integrates in its essence both causation and material constitution. The identification of those types is based on the very standards of mechanistic methodology, i.e., localization, decomposition and integration of parts and their operations. The methodology of mechanistic research is reflective of the metaphysics behind mechanistic frameworks and that metaphysics is inherently reductionistic. This means that despite mechanistic philosophy apparently offering a mixture of grounding (vertical "constituent" realization) and irreducibility (horizontal "causal" autonomy), higher-order property kinds cannot be explanatory subsumed nor reflected by the sum of mechanistic mereology. The main reason for this failure is that the grounding of mechanistic explanations cannot be separated from the requirement of that grounding being the expression of identity types.

2. The Mereological Character of Mechanisms

Mechanisms do exist in many natural and artificial contexts, and they constitute in part *epistemic tools* that help us comprehend the world. They also enhance our understanding of reality by representing not only empirically based heuristic methods but also by revealing the very structure of the world. Thus, the analysis and discovery of mechanisms represents an *ontological attitude* to the world, i.e., natural phenomena in the world are made up by mechanisms. Whatever phenomenon of interest we would like to analyze, it is then according to mechanistic philosophy a complex structure composed of units and their parts working together in the performance of activities at different levels of organization (Craver and Darden 2001). Mechanisms appear both between and within levels at different scales of organization in one and the same complex structure or system. Mechanisms compose or make up their referents or target phenomena. They explain the explanandum phenomenon by means of *major* or minor *sub-mechanisms.* A major mechanism would then be the kind of physical structure the organization of which encompasses or engulfs the key defining characteristics of a higher-level structure. The encompassing organization is instantiated by a constituent relationship, i.e., the sum total of the parts of the mechanism matches the totality of the phenomenon. A minor mechanism or sub-mechanism is a proper part of a major mechanism. The minor sub-mechanism is autonomous in terms of its unique or specific constitutive causal character, i.e., the minor mechanism defines a *certain* aspect of the more general phenomenon. Generally, a major mechanism is composed by minor mechanisms. Further, it is the defining features of the target phenomenon together with the level of analysis applied to it that makes the integration of a sub-mechanism into the wider hierarchy of the major mechanism relevant or not. In other words, a minor mechanism would, depending on the circumstances, be necessary and sufficient for the physical realization of a certain property and so exercise its constitutive causal effect on the wider phenomenon. Depending on the circumstances the minor mechanism is recruited or not into the functional organization of the wider mechanistic system. Those circumstances are reflective of the sensitivity of phenomenon to contextual and environmental effects. Moreover, the structural organization of the minor mechanism together with its defining functional roles is generalizable and applicable to distinct major mechanisms. Thus, once the relevant constitutive character of a minor mechanism is recognized as participatory in a plurality of different phenomena, the minor mechanism's causal profile is accordingly instantiable in a plurality of ways. Minor sub-mechanisms are then categorized according to their effective causal contribution to the formation of higher-order phenomena. In other words, one and the same sub-mechanism can be recognized as constitutive for a plurality of phenomena. As such it can be recruited into the workings of a major mechanism according to its explanatory relevance in the final analysis of the target phenomenon. Inversely, a major mechanism might recruit or even silence the effect of a minor mechanism in the general framework of the constitutive relationship. Thus, within the main mechanisms there exists sub-mechanisms behaving as constituents of entire mechanisms. This relationship between the effect of the sub-mechanisms on their major mechanistic counterparts and the effect of latter on the way the sub-mechanisms are expressed as part of the total defining systemic activity is the expression of a mereological relationship. Thus, mechanisms might vary in their functional and constituent characteristics but maintain their capacity to collectively explain the target phenomenon as the result of their participation in the grounding of a complex part-whole relationship. Moreover, mechanisms participating in mereological relations means that they explain the occurrence of a certain phenomenon by means of *physical realization* both causally and constitutive; mechanisms are physical entities. Causally speaking, the mechanisms instantiate and carry out the characteristic causal role description of the phenomenon. Constitutively speaking, mechanisms represent the very material occurrence of objects and their elements participating in the physical structuring and configuration of natural phenomena. Furthermore, mechanisms explain also by means of their *directionality.* This means that mechanisms in the context of their overall position in the "topology" of the comprising mereological relation are explanatory relevant either "vertically" or "horizontally." In a vertical explanatory understanding, major mechanisms and their recruited sub-mechanisms explain *synchronously.* This means that the very presence or manifestation of the mechanism in the topology of the entire system is sufficient for the higher-order phenomenon and its properties being instantiated. Here, "vertical" refers to the condition that the realizing relation is achieved through material synchronous constitution. There is then basically no need for a causal relationship to hold between the realizing mechanism and the higher-order object. To compare, any causation at the proper level of mechanisms is *internal* to the workings of single mechanisms. In other words, every sub-mechanism implements a certain functional role by means of integrating into its activity a causal role description. The execution and integration of the sub-mechanism's causal role in a wider upper-level mechanism is dependent on the overall *contextual characteristics* of the overarching mechanistic hierarchy. Thus, if a mechanism in general and its contributory sub-mechanisms in special explain something causally this is due to processes based at the *intra-level* causal structure of organization, i.e., the system's *horizontal directionality.* Importantly, causation as such is explanatory to the extent that objects at different levels of organization causally interact with each other. Composition as synchronously organizing mechanisms explain by way of their very physical occurrence not necessarily by their causal contribution. This difference is relevant when it comes to the kind of phenomena we want to mechanistically analyze. Thus, if mental properties like qualia are in principle *functionalizable* it means that their causal role description can be replicated in ways independent of the character of their implantation platform. Given this condition, the intra-level causal structure of the participating minor mechanisms is then also explanatory relevant. On the other hand, if qualia as an example of a mental category are not functionalizable, then causal explanation does not necessarily track them. Nevertheless, they can still be explained in the sense of their *occurrence as a matter of fact* or as the result of the very physical occurrence of the mechanical constituents composing them. That mechanisms do not primarily explain by casual structure can also be understood when considering that for example one and the same sub-mechanism can implement many *different* casual roles, some relevant other irrelevant, to the realization of one and the same higher-order phenomenon and its properties. This is so because, if mechanisms in general are intended to explain by means of *type-to-type identities,* then causation as described here cannot track the kind of stable identities required for those Types to obtain. The reason is twofold. First, type identities do not work through causal identifications when applied to non-functionalizable properties like qualia if functionalization understood here as psycho-physical functional instantiation fails. Second, the possibility of the same sub-mechanism instantiating one and the same higher-order property implies the multiple realization of the higher order phenomenon by its subvenient sub-mechanisms. Traditionally, *multiple realization* blocks type identities. Consequently, the concrete causal occurrence of a certain mechanism in the overall mereological hierarchy is not an absolute condition for the realization of the target phenomenon. Other alternative mechanisms *causally relevant* in the topological organization representing the target phenomenon can do the same work. It is then important to keep in mind that any causal effect related to a mechanism is an internally grounded process. Thus, the causal explanatory role of a mechanism is an *intra-level* local effect reflective of its specific structure and organization. The explanatory contribution made by a sub-mechanism is therefore a condition that can be isolated to the structural organization of the local sub-mechanism.

3. Explaining through physical constitution

Importantly, local sub-mechanisms are part of a major mechanism not causally but constitutively. This is so because any causal relevance on behalf of the sub-mechanism is an expression of the sub-mechanism's own self-contained *physical composition*. Thus, in terms of causal contribution or referring to casual powers the *causality* of the sub-mechanism is grounded on its own internal mereology. According to this picture, causation is physical constitution and constitution is mereology. Thus, the mereological relation is mainly explanatory in the physical constitutive sense, not necessarily explanatory in a causal sense. What the causal contribution of the sub-mechanisms add is to construe the flow of information relevant for the stability of the system as a process primarily localized to the internal composition of local structures. This has an important consequence for the analysis of the mind-body relation. Thus, for example, higher-order psychological phenomena like attitudes are mechanically explained as the constitutive "vertical aggregation" of a many sub-mechanical components acting together in a synchronous way. The attitudes *are* ontologically speaking this vertical aggregation, but we know or understand the behavioral dimension of the attitudes by the causal implementation of the involved sub-mechanisms, i.e., their horizontality. This means that a higher-order phenomena cannot be reductively understood as just as the total sum of causal effects realized by component mechanisms. As major mechanisms are "vertical" *constituents* of their co-related higher-order phenomena they are mereologically speaking not necessarily dependent on the *causal* contribution of sub-mechanisms for the realization of the phenomenon. There is no causal dependency relation holding between levels in a hierarchy of mechanisms but rather any dependency relation is only constitutive. To realize is to constitute not to cause. Thus, there is no inter-level *causal* constitution between the major mechanism and its sub-mechanisms. There is therefore at least a sense in which the target phenomenon is casually irreducible to the sum total of the causal role descriptions representing the whole set of sub-mechanisms. Mechanisms then metaphysically speaking constitute their associated phenomena; they do not cause them. The causal relevance of the sub-mechanisms is to perform a certain activity characteristic of its own local part-whole organization. At the same time, the very same sub-mechanism can implement many different causal roles and so there is no fixed or defining overarching causal role hallmark. Therefore, it is the very physical constituent-based occurrence of the sub-mechanisms as objects that matters for their contribution to the realization of the overall phenomenon. We could say that the *causality* of the mechanisms represents the *epistemology* part of mechanistic explanations while the *constitution* part is the *ontology* part of the explanations. This means that a phenomenon could in principle just be there so to speak without doing anything significant. Mechanistic models in general explain by tracking first and foremost the ontology not the epistemology. The epistemology guides the methodology, but the ontology grounds the phenomenon. Mechanistic models explain by tracking the ontology back to relations holding among *physical types*, thus mechanistic explanation is physical *type* constitution. Types are explanatory relevant as constructs tracking the existence of natural kinds. Thus, we have organisms composed by certain types of cells, certain types of tissues and organs as well as certain types of molecular pathways. As a result, mechanistic explanations of the mind and its causal powers are to a certain degree non-reductivist if mechanistic explanations playdown the role of causation and emphasize the mind's "vertical" constituent dimension. If causation plays any explanatory role in mechanistic philosophy, it is secondary and due to the *intra-level* or "within mechanisms" causation relevant to the identity of the mechanism. This is certainly most counterintuitive to many mechanistically oriented scientists. Especially, neuroscientist who view explanation as a causal relation holding *among* mechanisms at different levels.

4. Interventionism

According to mainstream mechanists, mechanistic explanation is in the first place an *inter-level causal* relation. I believe this reflects a confusion regarding the role that manipulation through *interventionism* plays when identifying the explanatory role of mechanisms behind the instantiation of a cognitive or mental function (Woodward 2007). Thus, the *inter-level-mechanist* believes that when intervening by means of physical intervention and manipulation on the function of a certain structure, the observed effects on the phenomenon are first and foremost *causally mediated* effects (Kubska and Kamiński 2021). Explanation then is a causal relation holding among mechanisms at different levels where the causal effects of one mechanism at one level explains the behavior of another mechanism at another level. Importantly, the sum effect of this interlevel causal relationship is also explanatory in terms of the type of mechanisms relevant for sustaining the integrity of the higher-order phenomenon. Thus, as an example in the neurosciences for this kind of inter-level causal explanation we can take the formation of spatial navigation memory maps (Farzanfar et al. 2023). The maps are the result of cellular and molecular mechanisms instantiating long-term potentiation and long-term- depression involved in the formation of spatial maps representations (Durand, Kovalchuk, and Konnerth 1996). An intervention on the cellular basis responsible for long-term potentiation and long-term depression reveals the causal contribution of the molecular level activity to the formation of the spatial orientation maps. Still, this is just one way that intervention and manipulation could be conceived as explanatory relevant. From an interventionist point of view regarding the character of causation, intervening means something being causally *symmetrical.* Thus, an intervention at the level of the phenomenon is also revelatory of the way that lower-level physical mechanisms contribute to the realization of the phenomenon. Interventionism represents a kind of causal *mutualism* in the framework of explanatory models. Still, this "mutualist" interpretation is limited because the intervention on the targeted mechanisms would also include more drastic interventions than just correlative manipulation. Thus, manipulation could refer to both invasive direct cell recordings or non-invasive techniques like transcranial magnetic stimulation (TMS)

(Leodori et al. 2022; Spampinato et al. 2023; Fregnac et al. 2009; Sparing and Mottaghy 2008). An invasive more drastic way of revealing the causal contribution of neuronal ensembles to the targeted function represents obliterating studies (Suzuki 2022; Qvist et al. 2018). Such studies reveal the importance and constitutive character of a certain component in the hierarchy of mechanisms once the targeted sub-mechanisms have been obliterated and removed (Gotzsche et al. 2022). Thus, adding or removing the occurrence of one or more sub-mechanism into the mereological hierarchy would be evidence also of their compositional in the realization of higher-order phenomena (Rolls 2021). One important consequence of understanding mechanisms as components is that the causal powers of the mental would be secured. In the total mereological relation among major mechanisms and their sub-mechanisms the entire part-whole system *is* the phenomenon represented by the total sum of relevant mechanisms. Sub-mechanisms are describable by their local intra-level causal profiles and by their capacity for *pluripotentiality*, i.e., meaning one and the same sub-mechanism being participatory in the realization of many different properties.

5. Mereological Irreducibility of Mind

Further, the mind is an example of a mereological relation embodied by the entire mind-body system. The causal profile of the individual mechanisms is then incorporated onto the causal profile of the entire system as an expression of its mereology (Juarrero 2015). Thus, putting things together, whatever causation there is, it is inherently distributed across the sum total of the system. Causation is then embodied at the intra-level of mechanisms; causal effects are primordially locally confined to the internal composition of the sub-mechanisms. There is then a kind of locally based causal closure that makes sub-mechanisms causally irreducible to each other. Given that sub-mechanisms are mereological components in the *entire* mind-body system, the mind is both causally effective and irreducible to just certain types of mechanisms in the part-whole topological configuration. This is so because, many different causal profiles can be instantiated by one and the same sub-mechanism and one and the same causal profile can be instantiated by many different sub-mechanisms. Also, one and the same overall mechanism and its component sub-mechanisms can be constituents participatory in many parts of the systems. Further, because of the inherent resilience and redundancy of many systems, especially biological ones, the entire system can remain functionally intact up to a certain degree with or without the presence of a specific main mechanism and its composing sub-mechanisms. The mind being an embodied systemic quality of the entire organism would then be irreducible to either the horizontal causal dimension of specific sub-mechanisms or to the vertical dimension of specific component mechanisms. As a result, we might say that the mind is an irreducible topological quality of the entire mind-body system. It is actively engaged with its surroundings by means of its embodiment in the realizing mereology of its constituent mechanisms. This has also an important consequence for the question whether mechanistic explanations for the mind are reductivist or not. Thus, the only way to secure a reductive turn on the mind in terms of reducing mechanisms is for the mechanists to insist on the occurrence of identity types. However, if the former observations are correct, this would not work because neither causally nor compositionally speaking can the mind be identified with mechanical types. As a result of the significance put on compositional mereology, we can discern a kind of *species chauvinistic* move. To be a certain organism is to be a certain conglomerate of certain types of microconstituents standing in compositional relation to each other both biologically and physico-chemically. In other words, an organism is a dog or a human only to the extent that *the right kind* or types of constituents are present in the mereology of the realizing biology. As we remember, according to this logic what matters is not so much causation being instantiated by different components but rather the very occurrence of the components themselves. Thus, in this sense mechanistically oriented explanations based on the mereological relation are in fact reductivist explanations but only to the extent that we insist on the reducing necessity of identity types. On the contrary, the very mereology of the system renders this requirement hollow. The mind then being an embodied systemic quality is both causally and compositionally irreducible to mechanisms.

6. Defining Mechanisms

As there are variations on the kind of properties that mechanists want to explain there are also variations of mechanisms. Furthermore, the very definition of what constitutes a mechanism varies. Thus, mechanisms are understood in a diversity of ways. Yet they hold certain qualities in common:

- “A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations.” (Glennan 2002, 344)
- “Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions.” (MDC 2000, 3)
- “A mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon” (Illari &Williamson 2012, 120)

Common to these definitions is the emphasis on entities being responsible for the existence of other entities. Mechanisms in general represent the instantiation of a behavior or the occurrence of a higher-level phenomenon by means of their constitutive parts or entities being systemically organized into realizing activities.

Mechanists can further be divided into three fractions or generations depending on how strictly they consider the constituent parts of their mechanistic explanatory models to be identical in relation to the explanandum phenomenon. Thus, *strict localization of function* mechanists, maintain a modular view of the mind-brain relationship viewing the mind as strongly correlated with the operation and activity of certain computationally encapsulated brain areas. Dynamical *systems-oriented* mechanists incorporate context sensitive and time-bounded aspects to the modular picture of the mind-brain continuum (Bechtel 2015, 2020). They usually refer to brain networks and circuits as the locus of identification in the brain for the execution and maintenance of a certain mental faculty. Mechanists referred to as *epigenetic-molecular* oriented mechanists take the environment and its effects on the molecular level of cellular organization as the locus of identification for cognition by means of referring to the role that molecular-genetic pathways play in the realization of higher-order mental properties (Bickle 2020; Tonegawa et al. 2015).

Even as there are similarities, there are also conceptual differences among these three examples of mechanistic generations. The differences refer mostly to the degree of *empirical resolution*; meaning the level of analysis at which mechanists carry out their investigations. Thus, mechanistic investigation encompasses descriptions of a higher-order phenomena all the way down from the contribution of specific brain areas and structures, through the way these structures represent patterns of brain-network activity, down to the lower level of molecular organization (Robertson and Cohen 2006). Mechanisms can then operate at *multiscale dimensions* and together contribute to the instantiation of the explanandum phenomenon. The level of empirical detail and resolution for the responsible mechanisms varies depending on the kind of mechanistic framework applied to the target explanandum. Thus, as the kind of explanations provided by mechanists moves from a more or less static view regarding the contribution of specific brain areas to a more dynamically oriented understanding of the mechanisms involved, so even changes their view towards a more pluralistic oriented interpretation of the phenomenon itself (Witherington 2014). In other words, the target phenomenon is considered the sum of different organizational scales operating together to enable the realization of the phenomenon (Khaluf et al. 2017). It must also be observed that not all sub-mechanisms necessarily contribute at the same time to the integrity of the target phenomenon as a particular entity. Rather, depending on the character of the phenomenon and often on the embedded dynamics in which it evolves, some sub-mechanisms might be activated while others deactivated, all depending on the circumstances (Deco et al. 2017). Thus, if we take episodic memory as an example, the character and quality of the episodic engrams is affected by the environmental circumstances where the experiences were acquired (Ramirez, Tonegawa, and Liu 2014). The environmental characteristics relevant to the memory acquisition process contribute to the characteristic content of the memory (Wang et al. 2024). These environmental effects activate different sub-mechanisms at different stages in both memory formation and retrieval (Ruiz et al. 2023). Meanwhile the identity of the target phenomenon, the episodic memory as such, remains specifically the same. Thus, the memory of granny's 80th birthday party remains basically the same semantic unit of information distinctive from alternative episodic engrams codified by alternative contexts and situations. Whatever the experiential content of the engram, subvenient neurophysiological mechanisms are active at different scales of physical organization. These scales involve time-bounded activity patterns covering different neuronal structures and network connectivity patterns supporting the distribution of separate cognitive functions (Buzsáki and Vöröslakos 2023, Buzsáki and Tingley 2023). One and the same cognitive function can then be mechanistically realized at the intersection of different scales. The scales range from network-based topology configurations referring to *structural connectome* issues to the functional pattern distribution of nerve cells firings referring to *information processing* issues (Sporns 2013). Therefore, differences exist at the level of the responsible brain-based mechanisms emphasizing both structure and stability as well as dynamism and activity. These constitutive differences at the level of empirical resolution affects also the kind of methodologies mechanists use in their analysis. Thus, while a mechanistic locationalist emphasizes the contribution of a certain brain area to a given psychological phenomenon, an epigenetic-molecular mechanist refers to the realization of the same phenomenon by means of molecular processes and pathways. Still, the three generations of mechanists share an implicit reductionist attitude in the sense that *there is* a basic computation instantiated by specific brain types. This computation, sometimes called a metacomputation, is carried out by a determined brain area or a specific molecular pathway performing a dynamical operation responsible for the integrity of the phenomenon (Piccinini 2010).

7. New Wave Reductionism

Further, all generations combine pragmatically into the framework of *New Wave Reductionism* which is essentially a localist enterprise oriented towards the establishing of type-to-type identity relations between brain mechanisms and mind (Richardson 1999; Bickle 1995). In addition, mechanists assume their models to be explanatory only to the extent that those models track the ontic structure of the world, i.e., the existence of distinguishable parts and objects responsible for the instantiation of observable phenomena and their behaviors. In other words, the models represent the *actual* physical conditions of realization for the phenomenon under investigation. Mechanists start by targeting a suitable phenomenon of investigation like memory consolidation or object pattern recognition. Based on their physicalist commitments and on the level of analysis being applied they proceed further through the following steps and strategies: decomposition of the phenomenon into its constitutive parts, i.e., the neurologically based properties, identification of the activities that those parts engage in, and the creation of a model that integrates the way those activities organize to form the targeted higher-order phenomenon. The decomposition aspect is essential. The identification of the constitutive parts is guided by a metaphysics and methodology reflective of things being made up and nested into each other. Once this strategy has been completed, the type of mechanism mainly involved in the process of decomposition is regarded as *synonymous with* the phenomenon itself. As a result, the explanandum phenomenon has also become *localized.* A crucial requirement for the process of decomposition to work is the following: the mechanisms must be concrete and real entities. In other words, the mechanisms answer to the ontic constraint as stated by Craver (2019), i.e., whatever the units of organization might be they must correspond to *real* things in the world. Neurologically speaking this means the localization of *concrete areas and pathways* in the brain responsible for the instantiation of the target cognitive function. Mechanisms being ontic means further the *reification* and *concretization* of the relevant parts and structures constitutive of the very phenomenon. Mechanistically oriented researchers in the neurosciences that conform to the former strategy are regarded as *classical mechanists.* In other words, they implement and develop the strategy of strict localization of function in line with the ontic constraint. They also search to identify *what* and *where* in the brain the different steps are specifically realized that correspond to the instantiation of a given cognitive function. Important in this procedure of mechanization of function is that the mechanisms as such *integrate* the different parts and steps of the process into a coherent and unified construct, i.e., the mechanism. Only then can mechanistic explanations have both epistemological as well as metaphysical weight. That is, the discovered major mechanism explains the phenomenon by being the phenomenon.

Another fraction of mechanistically oriented researchers, *the modern mechanists*, are ambiguous in their view on the relevance of the ontic constraint. They are very well aware of the *non-strictly localizable* and dynamically changing structure of many cognitive functions and the way their functional patterns are *temporally distributed* in the brain (McCaffrey 2023, 2015). This means that not only the *what* and *where* of the responsible entities such as brain areas bear the explanatory burden but also the "*when*" aspect of the time-bounded correlations between phenomenon and brain states is significant in terms of explanation (Love 2018). For the modern mechanist then any limitations based on the purely ontic relevance of the mechanisms representing real, recognizable and isolated objects is problematic. One reason for this is that mechanistic explanations as mentioned earlier, methodologically speaking rely most of the time on interventionist manipulation techniques to establish the mapping of brain structure to phenomenon. To manipulate and so intervene, the mechanists need to intervene on *concrete* objects of analysis. One caveat with this interventionist strategy is the way in which not only brain areas but also neuronal groups in general change and share functionality in the distributed network-like processing of information implemented by the brain (Ptak, Doganci and Bourgeois 2021). Thus, the more time-bounded and dynamical the behavior of the elements become the more difficult it becomes to *disentangle* them from their relational networks as so the more the difficult to intervene on, i.e., the intended object of intervention becomes less objectual and more difficult to formalize and operate on (van Gelder 1998; Eliasmith 1996; Chemero 2000). The issue here refers to the way mechanists are supposed to understand what it means to concretize and reify entities, i.e., their ontic constraint. Given the more time-bounded perspective on neuronal functionality and communication the character of the entities as such becomes more *malleable* and diffuse. As a result, the more difficult but not necessarily intractable it becomes to translate that dynamics into synchronous *static* models of behavior (Dewhurst 2018). Meanwhile, the classical mechanists have a stricter conception of how concrete and identifiable mechanisms are the same as the targeted phenomenon. For them, the dynamics of the system plays a less conclusive role compared to the coordination of parts by their participation in constitutive relations where structure determines function and causation. The modern mechanists are also committed to identification of mechanisms with the phenomenon, but they are open to questioning the requirement on identification based on strict localization of function. These different attitudes towards localization have consequences when it comes to the role that reduction plays in their mechanistic explanations. There seems to be two alternatives to understand the disagreement on how localization of mechanisms, either strict or distributed, could be compatible with the reduction of the phenomenon to relevant brain mechanisms. The alternatives can be summarized like this:

> **Alternative one**: the mechanisms are strictly reductive because they track identities in terms of type identities; *the explanandum phenomenon is the same as the corresponding brain area or brain state.*
>
> **Alternative two**: the mechanisms are not necessarily reductive because time-dependent and plasticity-based processes observed in the distributed dynamics of brain mechanisms blocks type identities. *Changes in functional networks over time, the brain dynamics, block the one-to-one reduction between the explanandum phenomenon and corelated brain state.*

8. Mechanistic Mereology

Mechanists of both sorts are also committed to the idea that phenomena showing systemic properties cover many levels of organization. Nevertheless, they disagree on the need for causally oriented explanations. Thus, for the dynamically oriented mechanists the actual realization of the phenomenon is determined by the mereological organization, i.e., the effects of the parts on the whole and vice versa. Thus, causation among levels and their objects is irrelevant for the relation between the whole and its parts. The part-whole relation is synchronously constitutive and not causal, i.e., any causation inherent to the system is derivative on the character of its constituent elements. Moreover, being a constitutive relation, *mechanistic mereology* explains the phenomenon in terms of *identities* holding between the relational and topological configurations among the whole and its parts. Thus, for a statue made of copper any causal operations relating parts to the whole at the level of physical organization are identical with the occurrence of the entire statue. Thus, any causal relations among parts of the statue are derivative from the fundamental copper structure. If the statue represents Don Quixote, then constitutively speaking the statue *is* Don Quixote. In a sense there cannot be any causal relations holding among parts of the figure. Thus, Don Quixote's nose is not *caused* by Don Quixote's arm. Rather any part of the statue is constituted by its participation in the form of the whole. Thus, Don Quixote's nose *is* identical with the topological composition of the whole statue system. Don Quixote and its copper statue is here reflective of a basically static relationship. Following this analogy, we may think that even organisms stand in this kind of constitutive relationship. Nevertheless, organisms not only are composed and configurationally located in space by their material structure but are also part of evolutionary and developmentally time-bounded processes. Still, mechanistic philosophy highlights the role that material composition plays in the *identification* of the parts involved. Thus, from a point of view where mechanisms represent reified parts, the phenomenon is identical with its realization mechanisms. The phenomenon must then be *the result of* a bottom-up process of analysis; from the most fundamental levels and their parts to the more derivative and complex ones. This process is based on recognizing the *relevant* level of physical organization, i.e., where in the system do we find the bottom-level of fundamental parts relevant for the integration of the phenomenon. Thus, fixing the relevant bottom-up level of analysis is necessary for the prospects of decomposition, localization and integration of activities to work. Given all these factors obtain, mechanistic philosophy assumes the *metaphysical completeness* of mechanistic explanations.

9. The Bottom Level

Without a relevant or common-ground level of analysis regarding the significant parts or entities from where to start decomposing and localizing, there is a risk that the mechanisms would become too broad and too diffuse and so metaphysically speaking vacuous and unreliable. This means that parts and their properties irrelevant for the explanation of the phenomenon could sneak in into the explanatory models. In other words, mechanists need to implement a principle or criterion of *demarcation and limitation* to establish what are relevant and reasonable boundaries for the identification of bottom-level mechanisms. Moreover, mechanists are physicalists. Thus, whatever that bottom level might be, it must be physical independently of the physical scale of organization referred to. Therefore, when studying the mind-brain relationship, neuroscientific mechanists often opt for the level of neurons and their physico-chemical properties as the relevant bottom line of investigation. As physicalists, they agree on that fundamentally there is only one relevant level of analysis for the mind-brain relationship to start with: the neuronal one; neurons are physical. At least, this should be the starting point from which mechanistic explanations on the brain-mind relationship should originate. Additionally, this way of fixating the relevant level of analysis would not necessarily exclude the possible contribution of *more basic* physical levels of organization. Thus, the *microphysics* of the brain is a contributory level of analysis for the understanding of brain structure and function. The leading metaphysical idea common to mechanistic philosophy is that we can follow the rabbit hole until we reach the very fundamental level from which the mind bottoms out. Here is where the reductive endeavors of the mechanists in general are exposed because they are interested in finding out that fundamental level of physical organization out of which higher-order properties arise. For the mechanistic neuroscientist this means *where* in the physical organization of neuronal components the sum and integration of those components gives rise to the mind. It is in their search for unity and fundamental organization that they share metaphysical commitments with the new wave reductionists (Bickle 2006; Bickle, De Sousa, and Silva 2022). New wave reductionists assume basically that there is only *one* relevant level of reality from which true explanations can be derived: one all-encompassing, undivided physical level of reality (Bickle, 2007; Hemmo & Shenker, 2022). Every mental phenomenon then is *ultimately* identical to the way in which the physical components at this basic level are organized and behave. Thus, mechanists and new wave reductionists share a metaphysical commitment to reductionism in their search for basic identities. They base their reductivist commitment on a ‘flat view’ of physical realty, i.e., any property in the world is ultimately the expression of synchronous and non-causal physical processes originating at that *fundamental level*. Such level encompasses everything there is in the world. All kinds of higher-level macrophenomena like language and perception can be explained by their relation to that constitutive “flat world” level. Also, mechanists of all generations follow new wave reductionists in their view on how mechanistic mereology explains higher-order properties like mental ones. Thus, mental properties are identical with physical mechanisms in a way that is basically synchronous and non-causal. As an expression of a part-whole relation the mechanization of brain function is characterized by the mereology of the system going all the way back to “flat world.” This means that the blueprint behind the organization of significant entities and their relational character at the bottom level is recognizable all the way up to more complex levels of organization. There is then a kind of bottom-up constitutive *mereological inheritance* ascending all the way up to higher-order levels of organization. In other words, the brain represents a mereological unit of organization and mereological inheritance represents a principle of organization. This principle of organization says that at whatever level of physical organization and scales we may encounter the existence of wholes, they are but the sum of their components. Furthermore, any systemic property we refer to as macro-based is the sum or relations holding among microscopic units. Neurons being the basic units of organization in the brain function as an example of physical structures following the mereological inheritance principle. Thus, neuronal structures inherit their organizational characteristics from basic mereological constitution, i.e., the functions of the brain are the expression of microscopic bottom-up constitution. The important question is then whether the inherited characteristic of micro based mereological organization is *continuous* in the sense that at no relevant level of organization higher up in the hierarchy, *emergent* or systemic properties might appear (van Hateren 2015). As an example, we can ask whether biological properties such as heredity and phenotype development are by their very biological character *different* from their more basic physical mereological underpinnings, i.e., their molecular-based genotype networks. A reductively minded neuroscience mechanist would probably agree with saying that at no relevant level of organization there is a break in the ontology of macro properties made up by their more basic parts. Therefore, fundamentally speaking no new properties other than those dictated by physics would organize in such a manner as to facilitate the realization of new functional configurations. In other words, biological properties are in no sense different from their physical constituents. More importantly, mechanists will agree that this mereological linearity will reveal basic identities at the bottom line of analysis being explanatory relevant for higher-level phenomena. In a sense, there is a fundamental level of *memory physics*, *language physics, love physics* and so on. The role of mechanistic explanations operating at the higher-level of the neuronal activity is to mimic the explanatory powers of fundamental physical science. Neuroscientific explanations based on the properties of neurons and their assemblies re-represent an ontological linearity starting with the physics of the brain. Mechanistic neuroscientists represent this mereological linearity by means of the kind of explanatory frameworks typically developed in their field. There are then metaphysically speaking *no ontological gaps* in the mereological understanding of the mind-brain system.

10. Special Sciences

In contrast, a special science researcher will disagree and say that there is no language or memory physics alone but that there is a *psychology* of memory and of language as well representing properties other than solely physical and neuroscientific ones. Thus, the meaning and semantics of language and the contextual embedment of memory contents represent properties not immediately readable from the physics of the brain. Furthermore, these properties are not unambiguously deducible from the syntax of the physical structures alone, but they are part and parcel of a level of organization other than just the neurological. Many higher-order properties are also open to contextual effects that even mechanistically speaking determine their specific qualities (Schauenburg et al. 2024). This is not to say that special scientists deny or undermine the relevance of psychological properties being physically and biologically grounded. It is only that those properties are not reducible to the former, basically because many of their intrinsic characteristics are either *multiply realizable* and/or pluripotential (Suzuki and Vanderhaeghen 2015).

Thus, independently of the mechanistic generation we refer to, there is a tension related to the understanding of what a mechanism is and on how to identify and delimit its structural and functional borders. Moreover, there is also a pressure between the way mechanists interpret and understand the explanatory power behind their mechanistic models. Thus, being aware of the dynamic character and behavior behind many organic processes, mechanists recognize the need and usefulness of dynamically oriented models and explanations. Still, many mechanists remain suspicious about dynamical models and their alleged explanatory powers. According to them the dynamics *describe* the behavior but to describe is not the same as to explain. According to the mechanists it is the synchronous presence and organization of the parts that matters for the development of real explanations in the neurosciences. Only by knowing what the relevant bottom-up neuronal structures are and how they relate to each other mechanistically can we develop explanatory models suitable for the neurologically based behavior of mental operations. Explanation consists then in the operation of the details, i.e., the objects and entities involved in the processes. The problem with this view is that the more the dynamics are recognized as integral elements to the nature of the phenomenon under analysis, the more the effects of intercorrelated patterns of activity play a role in the explanations. Moreover, the mechanisms themselves at the level of components behave dynamically (Bechtel 2019; Gilbert 2016). The more interconnected the parts, the more “nested” the character of the constituents. Mechanists seek for explanatory unity by highlighting the identity character of the components. Thus, at one level of analysis where constitution and parthood are primordial the relevant details are pretty much static and essentialist in character, i.e., unchangeable and fixed. At another level of analysis where the dynamics of the system and its complexity is considered, the more the malleable and irregular the identity of the parts. The tension then becomes obvious on the one hand between stable and recognizable elements and on the other hand partially concrete and dynamically malleable patterns of organization. Thus, the more dynamic, meaning the more distributed and flexible their identity, mechanisms as theoretical constructs explain in virtue of their participatory role in the activity flow. This is in line with the ontological commitment of mechanistic explanations; to explain is to discover the role that real objects of the world play in the making of natural phenomena. It is only that the reality of these components is now due to the dynamics of the system more flexible and relational. On the other hand, if the identity of the components becomes "unbounded" or muddied by the dynamics of the system then the more difficult to meet the requirements on localization, decomposition, identification and integration of objects into wholes. To solve this problem mechanistic philosophy would have to sacrifice its commitments to essential types and emphasize both the synchronous and diachronous character of the phenomenon. Thus, mechanistic explanations would need to take more seriously the time-bounded *diachronic evolution* of the phenomenon together with the *here and now* synchronous character of the components. Leaving out the diachronous "historicity" of the explanandum would from the point of view of many complex systems like biological ones seem almost impossible. For a living organism or an organ like the brain such a separation between its synchronous and diachronous dimensions would seem unbearable (Jablonka and Ginsburg 2022; Noble 2012). An organism as a living entity including all its embodied characteristics, both physiological and cognitive, is inseparable from the conjunction of both its synchronous and diachronous realization bases. One solution to this conundrum would be to sacrifice the ontic commitments of mechanistic philosophy and its requirements on "reification" and go full-blown "dynamic." Thus, mechanistic models would just pick up powerful *epistemic tools.* Mechanisms then pragmatically *describe* the phenomenon without any requirement on the identification of fixed identity types. The problem then is that mechanistic explanations would by means of becoming epistemic be *relativized* to the kind of descriptive background in which the dynamical processes are conceptually framed.; information flow diagrams, schemas, mathematical models and so on (Miłkowski and Hohol 2020). There is no more any "ontic power" or ontological weight behind the explanations and so they are no longer "mechanical" explanations at all.

11. Staying "Ontic"

The solution to this conflict would be to remain "ontic" and try to reify the relation between mechanism and dynamism. Mechanists would achieve this by insisting on that at the very ultimate level of physical analysis we still can describe the dynamics in term of identities. Moreover, we need to explore this possibility because of the requirement on *identification of parts* imposed on mechanistic explanations for them to make sense at all. If there are no identities in terms of clearly delimited parts, then mechanisms do not really explain, because it is the very identity of specific mechanistic entities that perform the work associated with the phenomenon. The problem with this solution is that at the very critical bottom-up physical level of analysis, there might be no more than *relational patterns of activity* and time-framed processes performing the work associated with the phenomenon (Davies and Gribbin 2007, Polkinghorne 2000). Thus, no matter what type of mechanist you are, appealing to the possibility of a physical bottom level of reality composed of delimited and unchanging parts won't work. A way to avoid this situation would be to say that the conflict between mechanistic and dynamical explanations is just linguistic. This is so because eventually only mechanisms and their parts are real. Dynamicists as well as mechanists refer in their explanations to properties the character of which represent natural structures and properties typical for *objects*. The dynamicist deals with objects the reality of which are mainly abstract while the mechanist deals with objects the reality of which is essentially physical. The mechanist either knows or intuitively recognize this condition as essential for explaining the matching of properties at different levels of organization to each other. The matching must refer to something being stable and recurrent enough to constitute an *object* of reality. Thus, my episodic memory of my granny's 80th birthday might be how embedded, dynamical and relational you want it to be but ultimately the memory is a natural kind the character of which picks up the presence of an undivided, recurrent and essentially "self-confined" *physical* property, the engram. It is this self-enclosed and delimited structure that makes the memory of granny's birthday party part of the world at all. If I want to explain the character of this memory in scientific terms then I do best according to the mechanists and new wave reductionists to frame it in an explanatory framework that involves a recognizable state of the world, i.e., an object. This object as a self-contained physical property is part of a memory-forming brain-based mechanism. Only in this way would any dynamical account of grandma's birthday party engram make sense. This way both mechanistic and dynamically oriented explanations are compatible. On the other hand, if natural kinds in general are not the sort of self-contained and indivisible properties the reductivists have in mind then the problem reappears. Thus, if even at the "flat world" level of Flat Physicalism preferred by many machinists as the appropriate metaphysical background, there really are no objects but rather continuously changing, coming and reordering multiscale *relational patterns*, then the very idea of objects and their parthood essential to the mechanists becomes rather hollow. Furthermore, if there really are no parts as self-contained objects then there are ontologically relevant mechanisms either. What really becomes a threat to the mechanist philosophy is then the possibility that the sort of natural kinds they refer to in their models of explanation bear no ontological status as *objects and things* and so cannot account for the discovery of identities. This is not to deny the existence of things being localizable and identical to the possession of certain recurrent properties in space and time. Rather, this is to recognize the possibility that what we take to be localizable natural kinds is but an epiphenomenon or at best an abstraction from a more basic relational space-time background (Austin 2016, 2020; Stengers 2008). If this is the correct interpretation of reality at the bottom level of physical organization then the principle of mereological inheritance would be broken. This in turn represents a bigger threat to the explanatory *completeness* of mechanistic models. This is mainly so because the braking down of mereological physical continuity would open for gaps in the ontology of the world. Those gaps might be filled by natural phenomena and their properties representing irreducibly autonomous units if organization at their own distinctive levels of organization. i.e., minds, species, economies. Also, natural kinds not being identical to types and their physical instantiations defies decomposition as the essential step in the final goal of mechanization of function which is the localization of identity types. Furthermore, a "relationalist" view on the character of natural kinds is compatible with the tenets of multiple realizability. Thus, given the case that one and the same higher-order phenomenon is stable enough to be realized by a multitude of different relational configurations and their properties, that phenomenon is multiply realized. A phenomenon can still be mechanistically realized by many different mechanisms. In that case, the realizing mechanistic units would because of their multiple realizability be *tokens* rather than types. In that case where a phenomenon is multiply realized by different mechanisms what we physically speaking at best can talk of are *tokened* physical properties instead of type-based identities.

12. Alternatives for the Mechanist

Nonetheless, because of their reductionist commitments, mechanists are compelled to reject these non-reductionist implications regarding the multiple realizability of mind (Baetu 2022). This is not to deny that the work of the mechanists in fact plays an important role in the understanding of how higher-order properties are physically instantiated by means of neuronal properties (Simons et al. 2017; Gouin et al. 2017). Nevertheless, mechanistic explanations cannot replace psychological explanations as such. The mechanists can *contribute* to the work done in the special sciences like psychology, but they must be prepared to give up their reductionist attitudes. They need to do this because of some unpalatable options reflecting the character of many higher-order systemic properties like psychological ones:

a) Higher-order properties have causal effects of their own. Higher-order properties are endowed with systemic effects recognizable at the lower level of their physical components. Thus, there *are* interlevel effects regarding properties located at different levels of organization; “the mind *moves* the body.” Mechanists must therefore reconsider to abandon their strong commitment to non-causal mereology blocking *interlevel* causation (Krickel, 2017).
b) Dynamical models of psychological functions do not only describe but explain. The behavior and function of cognitive agents is a dynamically grounded process best analyzed with the methods of systems theory and their mathematical tools. Mechanists must therefore at least be willing to accept the need for mechanistic models being

complemented by the virtues of time-dependent dynamical models; "the mind *happens* at different times."

c) Thus, they must also accept that many phenomena in nature are dynamical and that the proper level of *explanation* for those phenomena is at the quantitative abstract level; "the mind *abstracts from* the complexity of its environment."

d) They must also accept that cognitive phenomena are also neuronally distributed and plasticity-based processes cutting across levels and scales of organization in the brain. Distributed here means variability at the level of realization. The mind is thus a *multiscale realization phenomenon* based on a variety of physical bases available for its instantiation.

e) They must therefore be prepared to endorse multiple realizability as a serious metaphysical option.

f) Methodologically speaking, they must also be prepared to accept that multiple realizability represents a proper way of guiding research on the mind by means of emphasizing a *pluralistic* attitude to the brain-mind relationship.

The following is a final description of why mechanists should recognize their failure to explain properties endorsed by the special sciences in essentially mechanistic and reductivist terms (Fodor 2008).

13. Cognition as Adaptive Abstraction

Perception, memory systems and even language rely on both evolutionary and developmentally imposed constraining conditions. This is reflected on the way brains encode information at the neuronal level (Friston and Kiebel 2009; Bates 1994; Hardcastle 2001), i.e., the cognition. Thus, perception, memory and even language depend on limited physiological and computational powers. Analogically speaking, the cognitive work performed by the brain in the framework of those constraining conditions corresponds to the brain's software. The brain is also a physical object characterized by the performance of different mechanisms. Yet many of those mechanisms and the properties they instantiate are distributed and dynamic in character; the *where* and *when* of cognition. One essential function of the brain-cognition system is to *predict* adaptive behavior in terms of the recurrent re-evaluation of the information flow between perception and the neuronal systems encoding that information (Friston 2012; Mastrogiorgio 2022). The cognitive and perceptual work of the brain as prescriptive and anticipatory is dynamical in character (Hohwy 2020, 2021). This means perception and cognition are time-bounded processes with embodied, embedded, and enactive characteristics (Feiten 2020; Froese 2015; Silberstein and Chemero 2013). These characteristics are constitutive of a *cognitive agent,* and they bring with them borders as for example, between the agent's body perception–and the agent's physical world. As such the existence of these borders involve a necessary perceptual and cognitive *abstraction* from the details in the environment where the agent dwells. Such abstractions are dynamical in character. To function, they need to track the *actual* position and goals of the agent in its trajectories across physical phase-space. Those trajectories reflect the time-dependent dimension of cognitive and perceptual processes (Pöppel 2009; Smith and Katz 1996). Thus, the mapping of relevant physiological brain mechanisms onto cognitive processes is also dynamical in character. Such dynamism should facilitate the implementation of *abstraction* as a general feature in the agent's cognitive behavior. In other words, the agent's brain is continuously calibrating its position and goals in reference to the agent's proximal and distal needs by means of forecasting anticipatory models of reality (Krupenye 2016). The reliability of these anticipatory models is continuously evaluated in terms of their relevant adaptability and survival fitness to the organism. Furthermore, *abstraction from the details* is a continuous process formed by interplay between the brain and its environment and as such it represents a level of analysis proper to the macroscopic needs of the organism. Abstracting away from irrelevant details is adaptive and meaningful for a cognitive agent. There is no need for an agent to be aware of the detailed computational mechanisms and processes of its own brain; a kind of brain-based computational impenetrability. Cognition and brain form an *adaptive axis* of information processing, a *brain-cognition axis*, that is constrained, among other things, by the access and flow of energy available to the system (Pepperell 2018). What makes the abstractions at the level of the brain-cognition axis truly relevant for the needs of the organism is their capacity to enable the agent to react properly by making use of context sensitive brain-based networks representing the agent's cognitive, social and physical environment (Hovhannisyan and Vervaeke 2022). Thus, the brain forms abstractions that encode knowledge from its current environmental niche both *synchronously*–here and now–as well as *diachronously*, i.e., from past experiences. In this way it forms possible time-based prognoses and ways of action into the future preparing the organism in advance for possible outcomes; the *when* of cognition (Leuridan and Lodewyckx 2021). Mechanistic explanations in seeking to identify the agent's cognitive strategies with strict mechanisms and properties localized to specific brain structures- the *where* of cognition- fail to fully explain and cover the adaptive value and character behind "in time grounded" brain abstractions. The abstractions represent the dynamics of organismal networks the existence of which is "coarse," yet adaptive for the needs of the agent. The modelling of those networks, and the abstractions based on them, requires that we view the agent in its integrity as an adaptive dynamical system. The existence of *coarse cognition* and the kind of abstractions based on them represent the *how* and *why* of the brain-cognition axis. The brain is the carrier of macrolevel representations implemented by a macrolevel-organized creature whose adaptive needs have developed in response to reliable environmental signals exerting developmental pressure both on its genotype and phenotype (Mazzocchi 2008, 2016). Brains evolved under such conditions need to respond quickly and effectively to the ongoing communication between agent and its physical environment. This is achieved by the acquisition of a biologically based general adaptive capacity; the instantiation of *survival interface structures* both mentally and neuronally (Hoffman 2019). These brain-based survival interface structures are neuronally embodied circuits actively representing the world. They work without the need for explicitly detailed-loaded computations. These neuronally embodied circuits process *enough* information in a way just sufficiently detailed for promoting the subject's main goal, survival. That is, both the circuits and the informational structures they realized are *coarse* biological adaptations. This is how the brain systemically interacts with its world and why that interaction must for survival reasons be coarse but effective enough to enable the development of fast anticipatory mental and cognitive structures (Kano 2019). This is also why an organism's psychology is coarse because it lives in a world of macro dimensions involving both other agents and material macroscopic objects (Falikman 2023). The macro character of the world as perceived by the agent exercises an evolutionary pressure on its nervous system to develop the right kind of mental and computational representations. At the physical implementation level, the corresponding neuronal realization bases must also keep the same pace as the computational structures. Thus, they must also accommodate to the coarse character of the organismic needs. As a result, brains develop neuronal networks configurations to implement their ability to develop parallel distributed computations. These networks represent dynamical activation patterns specialized for abstraction from details both at the level of their single neuronal units as well as at the level of their connective topology (Quiroga et al. 2005). The main message here is that the characterization of these neuronal patterns is determined by their malleability, processual properties, dynamical patterns of organization and behavior as well as their multiple realizability (Anderson 2010, 2016).

Conclusion

The mind-brain relationship can be characterized in terms of mechanistic constitution. Here the synergistic effect of neural mechanisms represented both by single cell as well as network-based processes is simultaneously constitutive. It means that brain structure and function work as parts of a mereological relation. Here the system's activity results from components acting in parallel at many scales of organization through the instantiating effects of grounding mechanisms and their substructures. This organizational topology as instantiated by brain networks constitutes the physical realizing base on which higher-order properties such as mental ones depend on (Wu 2021). Under the condition of there being a difference between the intra-causal activity of minor sub-mechanisms and the interlevel coordinated activity of a main mechanisms, there is a possibility that the mind as a *systemic property* is even mechanistically speaking irreducible (Krickel 2018). This would symbolize the explanatory power behind mechanistic philosophy which assumes the *concrete* reality of mechanisms.

Mechanistic philosophy relies on three conditions. One being that we have a coherent common ground of analysis for the understanding of what it means to be a mechanism. The second refers to a methodological concern where decomposition and integration of *objects* and their relations into structural and functional complexes realize the characteristics of the explanandum phenomenon. The third condition refers to the possibility of *localizing* the phenomenon qua mechanism through the establishing of *type identities* representing the structural and functional complexes of mechanisms. Were all these conditions fulfilled it would imply the in-principle completeness of mechanistic explanations and the actual *mechanization* of the mind. This would be in line with the commitments of mechanistic philosophy. The analysis conducted in this work contends that hardly all these conditions are really met. Mechanists have a difficult time trying out an unambiguous definition of what it means to be a mechanism specially when considering the ontic constraint imposed on the *reification* of constituent parts. As we have seen ontological commitments behind mechanistic philosophy rely on an interpretation of natural kinds as basically *essentialist* objects. This is not the only possibility of conceiving of natural kinds even in the context of mechanistic explanations. A *process-oriented* view based on a *relationalist* grounding of natural kinds involves an alternative metaphysical view. This has effects on the role that neuronal structures and their components play in the realization of mental properties. We have also seen that decomposition and integration of components into the mereology responsible for the realization of higher-order properties requires having a clear-cut boundary for the implementation of *relevant* mechanisms. Such a boundary is in many cases the result of an *ad-hoc* decision based on the theoretical and methodological mechanist background of the researcher. Thus, there is no such epistemically independent fundamental level out of which bottom-out mechanistic explanations in general. Finally, many higher-order mental properties are integrated into contextually sensitive frameworks for the guiding of behavior. The social dependency of many intentional mental properties reflects their evolutionary roots as well as intrinsic adaptive value (Grosse 2010; Tomasello 2023). The synchronous and diachronous characteristics of *predictive models* as implemented by the brain guide behavior both probabilistic and intentionally (Friston 2010).

Such models are oriented towards a balance between costs and benefits with the goal of optimizing the organism's behavior. An important feature of this optimization is the development of both coarse mental and neuronal structures specialized in the abstraction from the details. The brain's abstracting mechanisms realized by large scale neuronal topologies and assemblies inform the organism about the relevant number of details necessary to compute the conditions of its macro-based environment. Such abstracting characteristics even when grounded on physical brain properties are not defined by one-to-one *type identities*. Rather, the *functionally enough* abstracting capacity of brains is based in the contextual multiscale properties of brain networks responsible for the *multiple realization* of mental states (Wu 2021). We recognize the irreducibility and context dependent character of the mind in the functional malleability of our attitudes, beliefs and intentions (Liu 2014). Mechanistic philosophy needs to address all these points either by reconsidering its metaphysical or methodological axioms or by way of reinterpreting the completeness of mechanistic explanations. The emphasizes in mechanistic explanations of the mind remains on the establishing of type identities between mental properties and neurophysiological brain states. Thus, mechanistic explanations are inherently reductionistic. On the other hand, much speaks against such a reductive view of the mind even when considering its physical underpinnings. The irreducibility of mental states and their dependency on physical states opens for a *non-reductive physicalist* metaphysics of mind. The result of this analysis leads to the conclusion that mechanistic explanations remain *reductive* about the mind.

References

Anderson, Michael L. 2007. "Massive Redeployment, Exaptation, and the Functional Integration of Cognitive Operations." *Synthese*, 159(3): 329–345.

Anderson, Michael L. 2010. "Neural Reuse: A Fundamental Organizational Principle of the Brain". *Behavioral and Brain Sciences*, 33(4): 245–266.

Anderson, Michael L. 2016. "Neural Reuse in the Organization and Development of the Brain." *Developmental Medicine and Child Neurology*, 58: 3–6. https://doi.org/10.1111/dmcn.13039

Austin, Christopher J. 2016. "The Ontology of Organisms: Mechanistic Modules or Patterned Processes?" *Biology & Philosophy,* 31(5): 639–662.

Austin, Christopher J. 2020. "Organisms, Activity, and Being: On the Substance of Process Ontology." *European Journal for Philosophy of Science,* 10(2): 1–21.

Baetu, Tudor M. 2022. "A Mechanistic Guide to Reductive Physicalism." *European Journal for Philosophy of Science,* 12(4): 1–26.

Bassett, Danielle and Olaf Sporns. 2017. "Network Neuroscience." *Nature Neuroscience,* 20(3): 353–364. https://doi.org/10.1038/nn.4502

Bassett, Danielle, Perry Zurn, and Joshua I. Gold. 2018. "On the Nature and Use of Models in Network Neuroscience." *Nature Reviews Neuroscience,* 19(9): 566–578. https://doi.org/10.1038/s41583–018–0038–8

Bates, Elisabeth. 1994. "Modularity, Domain Specificity and the Development of Language." *Discussions in Neuroscience,* 10(1–2): 136–149.

Bechtel, William. 2015. "Circadian Rhythms and Mood Disorders: Are the Phenomena and Mechanisms Causally Related?." *Frontiers in Psychiatry,* 6: 118–128.

Bechtel, William. 2019. "From Parts to Mechanisms: Research Heuristics For Addressing Heterogeneity in Cancer Genetics." *History and Philosophy of the Life Sciences,* 41(3): 1–23.

Bechtel, William. 2020. "Hierarchy and Levels: Analysing Networks to Study Mechanisms in Molecular Biology." *Philosophical Transactions of the Royal Society B–Biological Sciences,* 375: 1796–1805. http://dx.doi.org/10.1098/rstb.2019.0320

Beckage, Brian, et al. 2013. "More Complex Complexity: Exploring the Nature of Computational Irreducibility Across Physical, Biological, and Human Social Systems." In *Irreducibility and Computational Equivalence: 10 Years After Wolfram's A New Kind of Science*, Edited by Hector Zenil, 79–88. Berlin, Heidelberg: Springer Berlin Heidelberg.

Bickle, John. 1995. "Psychoneural Reduction of the Genuinely Cognitive – Some Accomplished Facts." *Philosophical Psychology,* 8(3): 265–285.

Bickle, John. 2006. "Reducing Mind to Molecular Pathways: Explicating the Reductionism Implicit in Current Cellular and Molecular Neuroscience." *Synthese,* 151(3): 411–434. https://doi.org/10.1007/s11229–006–9015–2

Bickle, John. 2020. "Laser Lights and Designer Drugs: New Techniques for Descending Levels of Mechanisms "In a Single Bound"?." *Topics in Cognitive Science,* 12(4): 1241–1256. https://doi.org/10.1111/tops.12452

Bickle, John, and André F. De Sousa, and Alcino J. Silva. 2022. "New Research Tools Suggest a "Levels–less" Image of the Behaving Organism and Dissolution of the Reduction vs. Anti–Reduction Dispute." *Frontiers in Psychology,* 13:1–12. https://doi.org/ARTN 99031610.3389/fpsyg.2022.990316

Buzsáki, György, and Mihály Vöröslakos. 2023. "Brain Rhythms Have Come of Age." *Neuron,* 111(7): 922–926.

Buzsáki, György, and David Tingley. 2023. "Cognition from the Body–Brain Partnership: Exaptation of Memory." *Annual Review of Neuroscience,* 46(1): 191–210. https://doi.org/10.1146/annurev–neuro–101222–110632

Cabral, Joana, et al. 2022. "Metastable Oscillatory Modes Emerge from Synchronization in the Brain Spacetime Connectome." *Communications Physics,* 5(1): 1–13.

Chemero, Anthony. 2000. "Anti–Representationalism and the Dynamical Stance." *Philosophy of Science,* 67(4): 625–647.

Craver, Carl. F. 2019. "Levels of Mechanisms: A Field Guide to the Hierarchical Structure of the World". In *The Routledge Companion to Philosophy of Psychology*, Edited by S.Robins, J. Symons and P. Calvo, 427–439. Routledge.

Craver, Carl. F. and Lindley Darden. 2001. "Discovering Mechanisms in Neurobiology". In *Theory and Method in the Neurosciences*, Edited by P. Machamer, R. Grush, and P. McLaughlin, 112–137. Pittsburgh: University of Pittsburgh Press.

Davies, Paul, and John Gribbin. 2007. *The Matter Myth: Dramatic Discoveries that Challenge Our Understanding of Physical Reality.* New York: Simon and Schuster.

Deco, Gustavo, et al. 2017. "The Dynamics of Resting Fluctuations in the Brain: Metastability and its Dynamical Cortical Core." *Scientific Reports,* 7(1): 1–14. https://doi.org/10.1038/s41598–017–03073–5

Durand, Guylaine M., Yury Kovalchuk, and Arthur Konnerth. 1996. "Long–Term Potentiation and Functional Synapse Induction in Developing Hippocampus." *Nature,* 381 (6577): 71–75.

Eliasmith, Chris. 1996. "The Third Contender: A Critical Examination of the Dynamicist Theory of Cognition." *Philosophical Psychology,* 9(4): 441–463.

Falikman, Maria. 2023. "Agency, Activity, and Biocybernetics: On The Evolution Of Agency. By Micheal Tomasello." *Mind Culture and Activity,* 30(1): 90–96. https://doi.org/10.1080/10749039.2023.2246947

Farzanfar, Delaram, et al. 2023. "From Cognitive Maps to Spatial Schemas." *Nature Reviews Neuroscience,* 24(2): 63–79. https://doi.org/10.1038/s41583–022–00655–9

Feiten, Tim. E. 2020. "Mind After Uexkull: A Foray Into the Worlds of Ecological Psychologists and Enactivists." *Frontiers in Psychology,* 11: 1–11.

Fregnac, Yves, et al. 2009. "Multiscale Functional Imaging: Reconstructing Network Dynamics from the Synaptic Echoes Recorded in a Single Visual Cortex Neuron." *Bulletin de L'academie Nationale de Medecine,* 193(4): 851–862.

Friston, Karl. J. 2010. "The Free–Energy Principle: a Unified Brain Theory?" *Nature Reviews Neuroscience,* 11(2): 127–138. https://doi.org/10.1038/nrn2787

Friston, Karl. J., and Stefan Kiebel. 2009. "Predictive Coding under the Free–Energy Principle." *Philosophical Transactions of the Royal Society B–Biological Sciences,* 364(1521): 1211–1221. https://doi.org/10.1098/rstb.2008.0300

Froese, Tom. 2015. "Enactive Neuroscience, the Direct Perception Hypothesis, and the Socially Extended Mind." *Behavioral and Brain Sciences*, 38: 22–24.

Gilbert, Scott. F. 2016. "Developmental Plasticity and Developmental Symbiosis: The Return of Eco–Devo." *Essays on Developmental Biology*, 116: 415–433. https://doi.org/10.1016/bs.ctdb.2015.12.006.

Gøtzsche, Casper R., et al. 2022. "Neuroglobin Deficiency Increases Seizure Susceptibility but Does Not Affect Basal Behavior in Mice." *Journal of Neuroscience Research,* 100(10): 1921–1932. https://doi.org/10.1002/jnr.25105

Gouin, Jean–Philippe, et al.2017. "Associations Among Oxytocin Receptor Gene (OXTR) DNA Methylation in Adulthood, Exposure to Early Life Adversity, and Childhood Trajectories of Anxiousness." *Scientific reports,* 7(1):1–14.

Grosse, Gerlind, et al.2010. "Infants Communicate in Order to Be Understood." *Developmental Psychology,* 46(6): 1710–1722.https://doi.org/10.1037/a0020727

Hardcastle, Valerie. 2001. "The Nature of Pain". In *Philosophy and The Neurosciences: A Reader*, Edited by P. Mandik, W. Bechtel, J. Mundale and R.S. Stufflebeam, 295–309.Blackwell Publishers Ltd.

Hermida, Margarida, and James Ladyman. 2025. "Physical Explanation and the Autonomy of Biology." *Philosophy of Science,* 92(4): 1–12. https://doi.org/10.1017/psa.2025.10115

Hoffman, Donald. 2019. *The Case Against Reality: Why Evolution Hid the Truth from Our Eyes.* WW Norton & Company.

Hohwy, Jakob. 2020. "New Directions in Predictive Processing." *Mind & Language,* 35(2): 209–223. https://doi.org/10.1111/mila.12281

Hohwy, Jakob. 2021. "Self–Supervision, Normativity and the Free Energy Principle." *Synthese,* 199(1): 29–53. https://doi.org/10.1007/s11229–020–02622–2

Hovhannisyan, Garri and John Vervaeke. 2022. "Enactivist Big Five Theory." *Phenomenology and the Cognitive Sciences,* 21(2): 341–375. https://doi.org/10.1007/s11097–021–09768–5

Jablonka, Eva and Simona Ginsburg. 2022. "Learning and the Evolution of Conscious Agents." *Biosemiotics,* 15 (3): 401–437. https://doi.org/10.1007/s12304–022–09501–y

Juarrero, Alicia. 2015. "What Does the Closure of Context–Sensitive Constraints Mean for Determinism, Autonomy, Self–Determination, and Agency?" *Progress in Biophysics & Molecular Biology,* 119 (3): 510–521. https://doi.org/10.1016/j.pbiomolbio.2015.08.007

Kaiser, Marie I., and Beate Krickel. 2017.”The Metaphysics of Constitutive Mechanistic Phenomena.” *The British Journal for the Philosophy of Science,* 68 (3): 745–779. https://doi.org/10.1093/bjps/axv058

Khaluf, Yara, et al. 2017. “Scale Invariance in Natural and Artificial Collective Systems: a Review.” *Journal of the Royal Society Interface,* 14(136): 1–20. https://doi.org/10.1098/rsif.2017.0662

Kano, Fumihiro, et al. 2019. “Great Apes Use Self–Experience to Anticipate an Agent’s Action in a False–Belief Test.” *Proceedings of the National Academy of Sciences,* 116(42): 20904–20909. https://doi.org/10.1073/pnas.1910095116

Krickel, Beate. 2018. “Saving the Mutual Manipulability Account of Constitutive Relevance.” *Studies in History and Philosophy of Science,* 68: 58–67. https://doi.org/10.1016/j.shpsa.2018.01.003

Krupenye, Christopher, et al. 2016. “Great Apes Anticipate that Other Individuals Will Act According to False Beliefs.” *Science,* 354(6308): 110–114. https://doi.org/10.1126/science.aaf8110

Kubska, Zuzanna Roma, and Jan Kamiński. 2021. “How Human Single–Neuron Recordings Can Help Us Understand Cognition: Insights from Memory Studies.” *Brain Sciences,* 11(4): 443–459. https://doi.org/10.3390/brainsci11040443

Leodori, Giorgio, et al. 2022. “The Effect of Stimulation Frequency on Transcranial Evoked Potentials.” *Translational Neuroscience,* 13(1): 211–217. https://doi.org/10.1515/tnsci–2022–0235

Leuridan, Bert, and Thomas Lodewyckx. 2021. “Diachronic Causal Constitutive Relations.” *Synthese,* 198 (9): 9035–9065. https://doi.org/10.1007/s11229–020–02616–0

Liu, Caimei, and Timothy C. Bates. 2014. “The Structure of Attributional Style: Cognitive Styles and Optimism–Pessimism Bias in the Attributional Style Questionnaire.” *Personality and Individual Differences,* 66: 79–85. https://doi.org/10.1016/j.paid.2014.03.022

Love, Scott A., et al. 2018. “Overlapping but Divergent Neural Correlates Underpinning Audiovisual Synchrony and Temporal Order Judgments.” *Frontiers in human neuroscience,* 12: 274–285. https://doi.org/10.3389/fnhum.2018.00274

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. “Thinking About Mechanisms.” *Philosophy of science,* 67(1): 1–25.

Mastrogiorgio, Antonio. 2022. “A Quantum Predictive Brain: Complementarity Between Top–Down Predictions and Bottom–Up Evidence.” *Frontiers in Psychology,* 13: 1–12.

Mazzocchi, Fulvio. 2008. “Complexity in Biology – Exceeding the Limits of Reductionism and Determinism Using Complexity Theory.” *Embo Reports,* 9(1): 10–14. https://doi.org/10.1038/sj.embor.7401147

Mazzocchi, Fulvio. 2016. “Complexity, Network Theory, and the Epistemological Issue.” *Kybernetes,* 45 (7): 1158–1170. https://doi.org/10.1108/K–05–2015–0125

McCaffrey, Joseph B. 2015. “The Brain’s Heterogeneous Functional Landscape.” *Philosophy of Science,* 82(5): 1010–1022.

McCaffrey, Joseph B. 2023. “Evolving Concepts of Functional Localization.” *Philosophy Compass,* 18 (5): 1–13. https://doi.org/10.1111/phc3.12914

Miłkowski, Marcin, and Mateusz Hohol. 2020. “Explanations in Cognitive Science: Unification Versus Pluralism.” *Synthese*, 199: 1–17.

Noble, Denis. 2012. “A theory of Biological Relativity: No Privileged Level of Causation.” *Interface Focus,* 2(1): 55–64. https://doi.org/10.1098/rsfs.2011.0067

Pepperell, Robert. 2018. “Consciousness as a Physical Process Caused by the Organization of Energy in the Brain.” *Frontiers in Psychology,* 9: 1–11.

Piccinini, Gualtiero. 2010. “The Mind as Neural Software? Understanding Functionalism, Computationalism, and Computational Functionalism.” *Philosophy and Phenomenological Research,* 81(2): 269–311. https://doi.org/10.1111/j.1933–1592.2010.00356.x

Polkinghorne, John. 2000. “The Nature of Physical Reality”. *Zygon,* 35 (4), 927–940.

Pöppel, Ernst. 2009. “Pre–Semantically Defined Temporal Windows for Cognitive Processing”. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 364 (1525): 1887–1896.

Quiroga, R. Quian, et al. 2005. “Invariant Visual Representation by Single Neurons in the Human Brain.” *Nature,* 435(7045): 1102–1107.

Qvist, Per, et al. 2018. “Brain Volumetric Alterations Accompanied With Loss of Striatal Medium–Sized Spiny Neurons and Cortical Parvalbumin Expressing Interneurons in Brd1+/− Mice.” *Scientific Reports,* 8(1): 1–12.

Ramirez, Steve, Susumu Tonegawa, and Xu Liu. 2014. “Identification and Optogenetic Manipulation of Memory Engrams in the Hippocampus.” *Frontiers in Behavioral Neuroscience*, 7: 76200–76209.

Richardson, Robert C. 1999. “Cognitive Science and Neuroscience: New Wave Reductionism.” *Philosophical Psychology,* 12(3): 297–307.

Robertson, Edwin M., and Daniel A. Cohen. 2006. “Understanding Consolidation Through the Architecture of Memories.” *The Neuroscientist,* 12(3): 261–271. https://doi.org/10.1177/1073858406287935

Rolls, Edmund T. 2021. “On Pattern Separation in the Primate, Including Human, Hippocampus.” *Trends in Cognitive Sciences,* 25(11): 920–922. https://doi.org/10.1016/j.tics.2021.07.004

Romero, Felipe. 2015. “Why There Isn’t Inter–Level Causation in Mechanisms.” *Synthese,* 192(11): 3731–3755.

Ruiz, Sara Arciniegas, et al. 2023. “Contextual Fear Response is Modulated by M–type K+ Channels and Is Associated with Subtle Structural Changes of the Axon Initial Segment in Hippocampal GABAergic Neurons.” *AIMS Neuroscience,* 10(1): 33–51. https://doi.org/10.3934/Neuroscience.2023003

Schauenburg, Gesche, et al. 2024. “Conflict Detection in Language Processing: Using Affect Control Theory to Predict Neural Correlates of Affective Incongruency in Social Interactions.” *Kolner Zeitschrift Fur Soziologie Und Sozialpsychologie*, 76(3): 603–625. https://doi.org/10.1007/s11577–024–00961–3

Silberstein, Michael, and Anthony Chemero. 2013. “Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences.” *Philosophy of Science,* 80(5): 958–970. https://doi.org/10.1086/674533

Simons, Ronald L., et al. 2017. “Methylation of The Oxytocin Receptor Gene Mediates the Effect of Adversity on Negative Schemas and Depression.” *Development and Psychopathology,* 29(3): 725–736. https://doi.org/10.1017/S0954579416000420

Smith, Linda B., and Donald B. Katz. 1996. “Activity–Dependent Processes in Perceptual and Cognitive Development.”. In *Perceptual and Cognitive Development*, Edited by R. Gelman and T. Kit–Fong, 413–445. Academic Press. https://doi.org/10.1016/B978–012279660–9/50030–0

Spampinato, Danny Adrian, et al. 2023. “Motor Potentials Evoked by Transcranial Magnetic Stimulation: Interpreting a Simple Measure of a Complex System.” *The Journal of Physiology,* 601(14): 2827–2851. https://doi.org/10.1113/Jp281885

Sparing, Roland, and Felix M. Mottaghy. 2008. “Noninvasive Brain Stimulation with Transcranial Magnetic or Direct Current Stimulation (TMS/tDCS)—From Insights into Human Memory to Therapy of Its Dysfunction.” *Methods,* 44(4): 329–337. https://doi.org/10.1016/j.ymeth.2007.02.001

Sperry, Roger. W. 1980. “Mind–Brain Interaction: Mentalism, Yes; Dualism, No.” *Neuroscience,* 5(2): 195–206. https://doi.org/10.1016/0306-4522(80)90098-6

Sperry, Roger. W. 1991. “In Defense of Mentalism and Emergent Interaction.” *Journal of Mind and Behavior,* 12(2): 221–245.

Sporns, Olaf. 2013. “Network Attributes for Segregation and Integration in the Human Brain.” *Current Opinion in Neurobiology*, 23(2): 162–171. https://doi.org/10.1016/j.conb.2012.11.015

Stengers, Isabelle. 2008. “A Constructivist Reading of Process and Reality.” *Theory Culture & Society,* 25 (4): 91–110. https://doi.org/10.1177/0263276408091985

Suzuki, Ikuo K., and Pierre Vanderhaeghen. 2015. “Is this a Brain which I See before Me? Modeling Human Neural Development with Pluripotent Stem Cells.” *Development*, 142(18): 3138–3150.

Suzuki, Hidenori. 2022. “Letter to Irreversible Neuronal Damage Begins just after Aneurysm Rupture in Poor–Grade Subarachnoid Hemorrhage Patients.” *Translational Stroke Researc*h, 13(3):355–356. https://doi.org/10.1007/s12975–021–00954–w

Tomasello, Michael. 2023. “Social Cognition and Metacognition in Great Apes: a Theory.” *Animal Cognition,* 26(1): 25–35. https://doi.org/10.1007/s10071–022–01662–0

Tonegawa, Susumu, et al. 2015. “Memory Engram Cells Have Come of Age.” *Neuron,* 87(5): 918–931. https://doi.org/10.1016/j.neuron.2015.08.002

van Gelder, Tim. 1998. “The Dynamical Hypothesis in Cognitive Science.” *Behavioral and Brain Science*s, 21(5): 615–625.

van Hateren, J. H. 2015. “The Natural Emergence of (Bio)Semiosic Phenomena.” *Biosemiotics,* 8(3): 403–419. https://doi.org/10.1007/s12304–015–9241–4

Wang, Yiyu, et al. 2024. “Memory Consolidation of Sequence Learning and Dynamic Adaptation During Wakefulness.” *Cerebral Cortex,* 34(2): 1–33. https://doi.org/10.1093/cercor/bhad507

Weiskopf, Daniel. A. 2011. “The Functional Unity of Special Science Kinds.” *British Journal for the Philosophy of Science,* 62(2): 233–258. https://doi.org/10.1093/bjps/axq026

Witherington, David. C. 2014. “Self–Organization and Explanatory Pluralism: Avoiding the Snares of Reductionism in Developmental Science.” *Research in Human Development,* 11(1): 22–36. https://doi.org/10.1080/15427609.2014.874763

Woodward, James. 2007. “Interventionist Theories of Causation in Psychological Perspective”. In *Causal Learning: Psychology, Philosophy, and Computation*, Edited by A. Gupnik and L. Schulz, 19–36. Oxford Academy Press. https://doi.org/10.1093/acprof:oso/9780195176803.003.0002

Wu, Zheng, and Bernhard A. Sabel. 2021. “Spacetime in the Brain: Rapid Brain Network Reorganization in Visual Processing and Recovery.” *Scientific Reports,* 11(1): 17940–17952.

Organon F 33 (1) 2026: 75–105 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2026.33103

RESEARCH ARTICLE

Interdisciplinary Learning by Bridging Theory and Literature: Exploring Feminist Discourse through Fiction

Henrieta Krupa*

Received: 26 August 2024 / Revised: 26 November 2025 / Accepted: 8 December 2025

Abstract: This article argues for the pedagogical value of using literary texts–specifically Virginia Woolf's *A Room of One's Own*–as an entryway into foundational concepts in feminist theory. It identifies three thematic clusters essential to feminist thought: (1) material conditions and gendered inequalities, (2) representation and discursive power, and (3) language, subjectivity, and gendered expression. Through close textual analysis, the article demonstrates how Woolf's narrative not only illuminates these themes but also renders abstract theoretical debates accessible to students encountering them for the first time. Integrating concrete classroom strategies, the study shows how Woolf's text can scaffold students' understanding of feminist concerns related to economic precarity, epistemic authority, gender construction, canonical exclusion, and the politics of language. Ultimately, the article contends that teaching feminist theory through literature enhances conceptual comprehension, strengthens critical and interpretive skills, and cultivates a nuanced awareness of how gender, power, and discourse intersect across both historical and contemporary contexts.

* Girne American University
https://orcid.org/0000-0003-3066-8813
✎ Girne American University, Faculty of Humanities, English Language and Literature Department, Universite Caddesi, P.K. 5, Karmi Campus, Karaoglanoglu, Kyrenia, Northern Cyprus
✉ henrietakrupa@gau.edu.tr

Keywords: feminist pedagogy; interdisciplinary learning; Virginia Woolf; feminist theory; literature in teaching.

1. Introduction

Teaching feminist theory frequently confronts a pedagogical challenge: many foundational concepts–discursive power, gender performativity, *écriture féminine*, canon formation, material oppression–may emerge as theoretically dense and sometimes too abstract for students encountering them for the first time. While these concepts form the backbone of twentieth-century Euro-American feminist criticism, they may often appear inaccessible when introduced solely through theoretical or philosophical texts. Literature, however, may offer an accessible, narrative-driven entry point that can help students grasp and contextualise these complexities. By its hybrid form combining fictional narration, reflective essay, and theoretical (philosophical) meditation, Virginia Woolf's *A Room of One's Own* proves to be particularly effective in bridging this gap between theory and accessibility.

Euro-American feminist criticism has developed through a variety of methodological approaches, each shaped by different historical and intellectual contexts. Elaine Showalter's (1981) influential overview in "Feminist Criticism in the Wilderness" captures this diversity by outlining a tripartite schema: British feminist criticism–largely Marxist in orientation–foregrounds oppression; French feminist criticism–shaped by psychoanalytic theory–emphasises repression; and American feminist criticism–often textual and literary in focus–centres on expression (186). Although Showalter's summary offers a concise historical snapshot, it seems to have significant limitations. The classification conflates intellectual approaches with geographical identities, obscures internal disagreements within each group, and fails to capture the profound cross-national exchanges that have shaped feminist theory. Simone de Beauvoir's (1949/1989) analysis of material oppression, for instance, aligns with themes Showalter (1981) labels *British*, while American (US) feminists such as Audre Lorde (1984), Kate Millett (1969/2000) and the Combahee River Collective (1977/1981) dismantle discursive constructions of gender in ways that exceed Showalter's (1981) *American* category. French theorists such as Luce Irigaray (1977/1985) and

Hélène Cixous (1976) differ markedly from Simone de Beauvoir (1949/1989), despite being placed under the same national label.

For these reasons, the article reframes Showalter's categories not as national taxonomies but as pedagogical themes that represent recurring questions in feminist thought across time and place. When taught through such a thematic lens, *A Room of One's Own* becomes a powerful text through which students can encounter three central clusters of feminist inquiry: 1) Material conditions and gendered inequality; 2) Representation, discourse, and canon formation; 3) Language, subjectivity, and gendered expression. This reframing allows students to engage critically with feminist ideas without inheriting the conceptual limitations of geographically defined categories. Woolf's novella is especially suited to introductory feminist pedagogy because it embeds many of key aspects of feminist theories within a compelling fictional scenario, providing a loop input for the mainstream Euro-American feminist criticism. The text is narrated by a fictional speaker–a crucial literary detail that invites students to question narrative authority, to interpret rhetorical strategies, and to reflect on the constructed nature of argumentation. Engaging with a fictional narrator also enables students to acquire theoretical knowledge through literary analysis, thereby modelling an interdisciplinary learning process that bridges literature and theory. In this way, the text introduces students to feminist concerns not abstractly but through concrete, contextualised examples that dramatise economic constraints, intellectual exclusion, representational distortions, and linguistic limitations.

Published in 1929, *A Room of One's Own* remains a foundational text in feminist circles, continually inspiring new scholarship and classroom applications. Not only does this work showcase Woolf's innovative narrative techniques–such as stream-of-consciousness, interior monologue, and self-reflexive commentary–but it also fundamentally resonates with major strands of Euro-American feminist thought that would be formalised decades later. Woolf's narrator's analysis of women's economic oppression and its impact on women's artistic expression prefigures materialist feminist arguments concerned with gender inequality; her reflections on women's absence from literary tradition speak directly to agitations that later shaped feminist criticism addressing issues of discourse, representation, and cannon formation; and her explorations of gendered language and creativity anticipate debates regarding language, subjectivity, and gendered expression as well as French feminist discussions of *écriture féminine*. Through its content and form alike, Woolf's text provides a rich textual foundation for introducing university students to these otherwise complex and often abstract theoretical frameworks.

The aim of this article is therefore twofold. The primary aim is pedagogical: to demonstrate how *A Room of One's Own* can be used strategically to teach major concepts in feminist theory across disciplines such as English Language and Literature, American Literature and Culture, English Language Teaching, Gender Studies, Women's Studies, and English Translation Studies. The secondary aim is literary: to illuminate how Woolf's text intersects with key strands of feminist thought and why this makes it an effective instructional tool. Through selective close readings and targeted classroom strategies, the present article offers a practical model for using Woolf's novella to scaffold students' comprehension of feminist theory and to cultivate interdisciplinary analytical skills. The paper proceeds by examining the three thematic clusters outlined above, pairing each with classroom strategies and concise theoretical connections. The goal is not to exhaust all possible theoretical intersections between Woolf's text and feminist theory, but to model how instructors can utilise selective, purposeful connections to enhance student learning as well as to illustrate how literary texts can serve as intellectually rigorous, engaging, and pedagogically effective gateways into feminist studies.

2. Fiction Enhancing the Teaching of Feminist Theory

Before turning to the thematic clusters, it is useful to outline why literature–especially hybrid literary-theoretical works such as Woolf's–supports the teaching of theories. The pedagogical value of teaching theory through literature draws on a long humanistic tradition that recognises narrative as a uniquely powerful medium for conveying complex ideas. As early as the Renaissance, Sir Philip Sidney in his famous *An Apology for Poetry (Or The Defence of Poesy)*, written around 1580 and published in 1595, argues that abstract philosophy often fails to reach the target audience because it usually communicates through generalisations and conceptual formulations that may feel remote or inaccessible. Poetry and fiction, by contrast, translate ethical and intellectual principles into vivid, memorable forms. As Sidney observes, philosophical knowledge "standeth so upon the abstract and general, that happy is the man who may understand him, and more happy that can apply what he doth understand" (Sidney 1973, 102). For Sidney, literature is therefore not a lesser form of knowledge but a more accessible one–an intellectual mode capable of engaging those who might struggle with purely abstract discourse. His claim that literature "can both teach and delight" (101), moving readers to understanding through concrete images, characters, and stories, anticipates modern pedagogical insights into how narrative scaffolds complex thought. The defence of literary meaning-making resonates strongly with the teaching of feminist theory today. Many feminist concepts–gender performativity, discursive power, structural oppression, or the politics of representation–can feel distant or inaccessible to students encountering theoretical language for the first time. Yet when these same ideas are embedded in narrative, metaphor, or fictional scenarios, they become easier not only to grasp but also to connect to lived experience. Texts such as Virginia Woolf's *A Room of One's Own*, with its interplay of argument and story, exemplify the pedagogical role Sidney envisioned. Woolf's novella offers philosophical insight through form that engages, delights, and illuminates by animating them through imagined histories, metaphorical scenes, rhetorical play, and a vividly constructed narrative voice. In doing so, her work models the very interpretive, reflective, and critical habits that feminist analysis demands. By embedding theoretical concerns in the texture of narrative–through story, character, imagery, irony, and voice–Woolf transforms complex feminist arguments into experiential knowledge. Literary texts thus operate as cognitive bridges, enabling students to approach abstract feminist concepts through the more intuitive logic of narrative, and without compromising analytical depth.

One of the primary reasons literature enhances the teaching of theory lies in its function as cognitive scaffolding for complex ideas: it provides concrete, embodied, and narrative-rich contexts through which students can grasp abstractions that might remain opaque in dense, purely theoretical prose (William and William 2003; Marek 2006; Cunningham 2024; Hansen

2023; Yazell et al. 2021). Literature, by contrast, grounds philosophical ideas in concrete detail. Woolf's novella accomplishes this by presenting hypothetical scenarios, imagined histories, and vividly staged scenes that render theoretical concerns more tangible. Her portrayal of a fictional narrator, her reconstruction of Judith Shakespeare's fate, or her contrasting depictions of Oxbridge meals embody feminist arguments about material inequality, patriarchal structures, and gendered exclusion. Through narrative, students witness theory in action. Instead of encountering concepts in isolation, they see how these ideas inform and shape lived realities. Fiction thus may become a pedagogical intermediary: it supports comprehension, provides memorable mental images, and allows students to process abstract ideas incrementally.

Closely connected to this is the historical entanglement between feminist theory and literary form. Feminist thinking has long emerged through literary modes of expression. Many feminist theorists–from Mary Wollstonecraft's *A Vindication of the Rights of Woman* (1792/1999) and Simone de Beauvoir's *The Second Sex* (1949/1989) to Hélène Cixous's "The Laugh of the Medusa" (1976)–blend argumentation with rhetorical flourish, autobiography, allegory, and metaphor. These works resist rigid disciplinary boundaries, demonstrating that feminist theory is not neatly divisible into *literary* and *philosophical* categories or *fiction* and *non-fiction* alike. Instead, it thrives through hybridity. Woolf's novella exemplifies this tradition: it merges fictional narration with social critique, historical analysis with speculative narrative, humour with earnest argument as well as fabrication with personal observation and experience. When students study feminist theory through such texts, they are not merely learning feminist concepts; they are being introduced to the very modes of expression through which feminist thought has historically evolved. The literary form itself becomes evidence of feminism's methodological innovation.

Additionally, literature fosters critical reading as a feminist practice, a skill fundamental to the discipline. Feminist inquiry demands attentiveness to silence, absence, and voice; to who is speaking and who is spoken for; to power structures encoded in language; and to representational norms and biases that shape cultural meaning. Close reading–understood as the careful, sustained analysis of textual detail–is therefore inherently aligned with feminist aims. Classic feminist theorists have long emphasised this connection: Adrienne Rich (1971) frames feminist reading as an *act of re-vision* that uncovers what has been suppressed or distorted; Sandra Gilbert and Susan Gubar (1979) demonstrate how attentive interpretation exposes gendered power within narrative structures; and Elaine Showalter (1985) and Toril Moi (1985) articulate close textual engagement as central to feminist critique. Recent scholarship continues and expands this tradition, highlighting close reading as a flexible, accountable feminist method attuned to voice, positionality, and interpretive responsibility (Holmes 2021; Lukić and Sánchez-Espinosa 2019; Kier-Byfield 2024). Woolf's text, for example, invites students to question narrative authority, trace ideological assumptions in male-authored history books, examine the symbolic significance of domestic and academic spaces, and interpret metaphors of confinement and possibility. Such practices cultivate the analytical habits that later allow students to engage more confidently with theoretical arguments about discourse, gender construction and representation, intersectionality, canon formation, and artistic production and distribution. Through literary analysis, students learn to recognise how power operates not only in institutions but also in language itself.

The interdisciplinarity of feminist methodology further underscores the value of literature in teaching feminist theory. Feminist inquiry traditionally draws from a wide range of fields–history, sociology, philosophy, linguistics, anthropology, psychology, and cultural studies (Jaggar 1983; Harding 1987; Haraway 1988; Lykke 2010; Hesse-Biber 2012). Literature provides a flexible platform that naturally accommodates these intersections, functioning as a site where narrative, theory, history, and cultural analysis converge (Hemmings 2011; Eagleton 1991). Woolf's novella, for instance, invites historical contextualisation (women's education in the early twentieth century), sociological analysis (class mobility, labour divisions), psychological exploration (interior consciousness, identity formation), philosophical reflection (epistemic authority), and linguistic inquiry (gendered language and the construction of subjectivity). Because literary texts contain narrative, imagery, historical reference, and philosophical speculation, they allow instructors to demonstrate the interdisciplinarity of feminist theory seamlessly. Students come to understand that feminist analysis is not an isolated intellectual activity but rather an integrative method capable of crossing disciplinary boundaries.

Another crucial pedagogical advantage of literature is its ability to engage and motivate students. Fictional scenarios, humour, irony, and vivid description often resonate more deeply with students than abstract theory, fostering cognitive and emotional involvement in the learning process (Freire 1986; Rosenblatt 1994; Egan 1997; Coles 1989). Woolf's narrator, with her conversational tone and self-reflexive wit, creates an intimate relationship with the reader, inviting curiosity and reflection. Students become invested in the story of Judith Shakespeare, the narrator's wanderings, or the absurd logic of male-authored treatises on women. These narrative elements make feminist ideas relatable and memorable. Instead of feeling overwhelmed by theoretical abstraction, students find themselves drawn into a reflective, imaginative space where learning becomes an active, pleasurable process. Literature's capacity to evoke empathy, provoke emotional responses, and stimulate interpretive engagement significantly enhances motivation–an essential factor in classroom learning.

These qualities illustrate why Woolf's *A Room of One's Own* in particular becomes an ideal pedagogical resource for introducing feminist concerns in an accessible yet intellectually rigorous ways. Its blend of storytelling and argumentation, its engagement with historical realities, its interrogation of discourse, and its stylistic experimentation all align with central concerns of feminist theory. The text both models and enacts feminist methodology: it critiques patriarchal institutions, analyses material conditions, exposes gendered structures of representation, and explores the relationship between language and subjectivity. At the same time, it demonstrates the interpretive skills–close reading, contextual analysis, critical inquiry–that students must develop to engage meaningfully with feminist scholarship. By providing theoretical insight within a literary form, Woolf's text offers students a way to approach feminist concepts not as abstract demands but as lived, embodied experiences. In this sense, literature does more than supplement the teaching of feminist theory; it transforms the learning process itself. It enables students to enter feminist debates through narrative familiarity, emotional resonance, and intellectual curiosity. Through works such as *A Room of One's Own*, students discover that feminist theory is woven into everyday life as well as language. Literature thus becomes both a pedagogical tool and a methodological bridge–one that leads students from textual experience to theoretical understanding with clarity, depth, and engagement.

3. Material Conditions and Gendered Inequality

Material conditions–economic independence, access to education, labour distribution, and institutional exclusion–are central concerns in feminist thought worldwide. Woolf's *A Room of One's Own* provides a compelling entry point into these themes through vivid narrative scenes that dramatise economic disparities and their effects on women's intellectual and creative potential. The text's pedagogical strength lies in its ability to concretise abstract feminist concepts, offering students a clear framework for understanding the material and structural constraints shaping women's lives.

3.1. Conditions of Creativity

Early in the text, Woolf's narrator famously asserts that "a woman must have money and a room of her own if she is to write" (Woolf 2004, 4). This line serves as a gateway into discussions about the economic restrictions historically faced by women, the relationship between financial independence and creative autonomy, and the gendered nature of space, privacy, labour, and institutional access.

Situating Woolf within the longer history of feminist thought enriches student understanding. Feminism's intellectual roots reach back to Enlightenment debates about social and political rights, which–though not explicitly feminist–laid the groundwork for later critiques of gender inequality. Mary Wollstonecraft's *A Vindication of the Rights of Woman* (1792/1999) is particularly relevant here, advocating for women's equal access to education. Woolf extends this argument more than a century later by demonstrating how limited educational privilege suppresses not only intellectual development but also artistic expression. Subsequent legal reforms such as the Married Women's Property Act (1882) and the Representation of the People Act (1918) further illustrate the slow, uneven expansion of women's civil rights, forming the historical backdrop against which Woolf exposes the economic disadvantages that shaped women's intellectual and creative lives.

Woolf's narrator, imagining herself addressing an audience of female scholars at Oxbridge, offers a sharply drawn contrast between the sumptuous luncheon provided to male academics and the meagre dinner served to women. These scenes provide a vivid visual metaphor for gendered educational inequality, illuminating the gendered nature of space and institutional wealth, and underscoring how institutional wealth structures opportunities for intellectual growth. These scenes in the text also form a natural bridge to Simone de Beauvoir's analysis of women's secondary status in *The Second Sex* (1949/1989), demonstrating how material deprivation and spatial exclusion reinforce women's subordinate position. Furthermore, the juxtaposition invites students to consider contemporary feminist concerns with intersectionality–how gender intersects with class, institutional privilege, and the politics of space.

3.1.1. Teaching Strategy: Guided Close Reading

Students can annotate the sensory details of the two meals scenes and discuss how institutional resources structures intellectual opportunity. The activity can then extend to identifying contemporary parallels–such as disparities in university funding or gendered patterns in workplace allocation. This approach offers a concrete illustration of a key Marxist feminist principle: material conditions shape consciousness, a principle echoed across thinkers from Mary Wollstonecraft *A Vindication of the Rights of Woman* (1792/1999) to Silvia Federici's *Caliban and the Witch: Women, the Body and Primitive* (2004).

3.2. Judith Shakespeare and the Structural Barriers to Genius

Woolf's thought experiment about Shakespeare's imagined sister, Judith, offers a powerful pedagogical tool for introducing students to structural inequality. Judith is endowed with the same innate talent as her brother, yet every aspect of her social environment–lack of schooling, limited mobility, parental expectations, sexual vulnerability, and exclusion from the theatre–conspires to extinguish her creative potential. Through Judith's story, the narrator demonstrates that women's historical absence from the literary canon is not the result of biological inferiority but of systemic restriction.

This narrative allows students to grasp a core feminist argument: creativity is shaped by social conditions, not natural aptitude alone. Woolf's portrayal of Judith's curtailed life resonates strongly with feminist thinkers who have analysed the material and institutional obstacles facing women artists. For example, in *Silences,* Tillie Olsen (1978) documents the silencing of working-class women writers, how working-class women, mothers, and women burdened by domestic labour were systematically prevented from developing their talents. Her focus on the *silenced* writer directly parallels Woolf's construction of Judith as a figure whose potential is crushed long before it can manifest. Silvia Federici (2004), in *Caliban and the Witch*, similarly argues that reproductive and domestic labour, imposed through patriarchal and capitalist structures, restricts women's access to creative and intellectual work. Federici's analysis helps students see how the constraints placed on Judith are not historical anomalies but symptoms of broader socio-economic systems. Judith's story thus becomes an accessible narrative gateway through which students can enter these broader theoretical debates about labour, gender, and creativity.

In addition, the exclusion of women artists can be extended beyond literature to other artistic fields. Linda Nochlin's (1971) landmark essay "Why Have There Been No Great Women Artists?" provides a parallel critique within art history, demonstrating how institutional, educational, and economic barriers–not lack of talent–explain the historical invisibility of women artists. Introducing Nochlin helps students recognise recurring feminist concerns across disciplines and artistic traditions. Consequently, the Judith episode may offer a productive entry point into contemporary feminist activism. Students can explore how Judith's fictional obstacles echo in the modern struggles around representation, publication, and recognition–issues addressed by groups such as the Guerrilla Girls, whose campaigns expose ongoing sexism, racism, and classism within the art world. These connections allow students to see Woolf's text not as a historical relic but as an early articulation of problems that persist today.

3.2.1. Teaching Strategy: Collaborative Reconstruction

In small groups, students reconstruct Judith's imagined life, identifying key structural barriers and mapping them onto present-day inequities. They then compare Judith's obstacles with those faced by marginalised writers today, making intersectional connections across class, race, sexuality, and nationality. This exercise encourages students to see Woolf's narrative as both historically grounded and urgently contemporary.

3.3. Intersection: Class, Gender, and Space

Woolf deepens her critique of material inequality by explicitly linking gender to class. The narrator observes that "genius like Shakespeare's is not born among labouring, uneducated, servile people" and asks how such genius could have emerged among women whose labour began "almost before they were out of the nursery" and who were constrained "by their parents" and "by all the powers of law and custom" (Woolf 2004, 56–57). In this reflection, women appear as a social class in their own right: structurally positioned, economically disadvantaged, and systematically denied the conditions necessary for intellectual development.

The concept of gendered space and class further illuminates this argument. Woolf's repeated emphasis on rooms, libraries, lecture halls, and dining spaces demonstrates how architecture itself can become a material manifestation of gender hierarchy. This theme connects readily to contemporary feminist debates–including those within British radical feminism–about women-only spaces, institutional belonging, and the politics of access.

3.3.1. Teaching Strategy: Mapping Gendered Space

Students brainstorm contemporary examples of gendered spaces–both exclusionary and protective–and analyse how such spaces shape gender identities and opportunities. This activity helps students recognise the spatial dimension of inequality and encourages them to consider the socio-political implications of who is allowed to occupy which spaces, when, and under what conditions.

3.4. Introducing Theoretical Connections

Woolf's narrative form provides an ideal opportunity to explore another dimension within feminist inquiry: the interdependence of artistic representation and lived material conditions. *A Room of One's Own* continually blurs generic boundaries between essay, fiction, satire, and lecture, thereby enacting the very argument it advances–that art does not stand apart from social reality but is shaped by it.

Directing students' attention to this formal experimentation might help them understand feminist theory's longstanding concern with genre, narrative voice, and the politics of representation. In *A Room of One's Own,* Woolf's narrator, for instance, imagines addressing a group of women scholars at Oxbridge, a scene that evolves from Woolf's earlier non-fiction essay entitled "Women and Fiction" (1929), which is based on the author's two lectures she delivered at Cambridge in 1928 (Rosenbaum ed. 1992). Introducing this context serves two pedagogical functions: it situates students within Woolf's real intellectual milieu, and it demonstrates how the novella transforms documented historical events into a fictionalised, self-reflexive narrative. The text's opening lines illustrate this transformation vividly. Beginning *in medias res*–"But, you may say, we asked you to speak about women and fiction–what has that got to do with a room of one's own?" (Woolf 2004, 3)–Woolf adapts a technique associated with classical epic while simultaneously critiquing the patriarchal literary inheritance from which that technique derives. The narrator's shifting identity–"call me Mary Beton, Mary Seton, Mary Carmichael or by any name you please" (5)–destabilises the assumption that the speaker is identical with the historical author. This deliberate ambiguity complicates distinctions between author and narrator, fact and fiction, lived experience and imaginative elaboration. As the narrative progresses and the fictional narrator gradually recedes, readers are left in productive uncertainty about who is speaking–an uncertainty that performs, rather than merely describes, a key Marxist-feminist claim: that artistic form is inseparable from the material and ideological structures that shape it.

Foregrounding this interplay between literary form and feminist argumentation enables instructors to introduce students to a broader constellation of feminist thinkers. Woolf's analyses resonate with Simone de

Beauvoir's (1949/ 1989) account of how social conditions shape women's possibilities; with bell hooks's (1982, 1984) critique of economic marginalisation and structural oppression; with Tillie Olsen's (1978) reflections on the silencing of women's creative labour; and with Silvia Federici's (2004) examination of reproductive labour under capitalism. These theorists need not be introduced exhaustively. Instead, selective pairing–using Woolf as a pivot–demonstrates how the text anticipates, echoes, or complicates major strands of feminist thought.

3.4.1. Teaching Strategy: Genre and Critical Discourse Analysis

Students examine passages where Woolf explicitly collapses the boundaries between fact and fiction–for example, the fictional reimagining of her 1928 Cambridge lectures or the narrator's refusal of a fixed identity. Through guided close reading, students analyse how Woolf's narrative strategies shape meaning and reflect on how form enacts a Marxist-feminist claim: that art is not an autonomous realm but one deeply entangled with the social world.

3.5. Addressing Critiques of Woolf

A rigorous pedagogical approach must also acknowledge that *A Room of One's Own* does not present a universal feminist perspective. While Woolf's insights into material inequality and women's exclusion remain influential, they are shaped by her own social position and by the historical circumstances in which she wrote. Introducing students to these critiques enriches their understanding of the text and deepens their engagement with the broader feminist conversation.

One prominent critique concerns Woolf's emphasis on money, private rooms, and access to education–conditions that reflect a distinctly bourgeois ideal of artistic production. Working-class women, colonised women, and Black women rarely possessed such resources, which raises critical questions about who is included in Woolf's imagined community of women writers. Furthermore, while Woolf foregrounds gender oppression, her analysis pays limited attention to the racial and imperial contexts that shaped women's lives in Britain and across its colonies. These omissions remind students that feminist theory must be understood in relation to its historical and social frameworks, not as a universally applicable model.

Engaging with these limitations does not diminish the value of Woolf's work; rather, it situates her arguments within ongoing debates about privilege, access, and intersectionality. Students learn to recognise how feminist texts both illuminate and obscure different dimensions of women's experiences, and how subsequent feminist thinkers–particularly Black, postcolonial, and working-class feminists–have expanded and challenged Woolf's claims.

3.5.1. Teaching Strategy: Reflective Writing Prompt

To foster critical engagement, instructors can ask students to respond to the following question: *Who is excluded from Woolf's vision of the woman writer, and why does this matter for feminist theory?* This exercise encourages students to reflect on the relationship between privilege and creativity, the role of material inequality in shaping artistic possibilities, and the importance of intersectional approaches in contemporary feminist scholarship.

4. Representation, Discourse, and Canon Formation

Representation and discursive power form a second major cluster of feminist thought that *A Room of One's Own* introduces with exceptional pedagogical richness. Woolf's British Museum episode, in which the narrator seeks authoritative knowledge about women, becomes a productive gateway into the politics of discourse, canon formation, and the ideological construction of gender. Through this scene–as well as through her discussions of literary stereotypes, narrative binaries, and the absence of a women's tradition–Woolf equips instructors with an accessible entry point into several strands of feminist theory, including discursive power (Beauvoir 1949/1989; Millett 1969/ 2000), performativity (Butler 1990; Butler 1993), feminist poetics (Gilbert and Gubar 1979), and *gynocriticism* (Kolodny 1980; Showalter 1988/1997).

4.1. The British Museum Scene as Gateway to Discursive Power

When Woolf's narrator visits the British Museum to "consult the learned authorities" on the nature of women, she finds shelves filled almost exclusively with books written by men–texts portraying women as inferior, irrational, morally weak, or inherently subordinate (Woolf 2004, 30). The narrator notes the striking uniformity of these male-authored accounts, which range from pseudo-scientific claims to religious condemnation, hostile opinion, and trivial anecdote.

This scene offers an accessible entry point into key feminist concerns: the discursive construction of gender, the gendering of epistemic authority**,** the historical male dominance in knowledge production, and the misogyny embedded in ostensibly *objective* scholarship**.** The British Museum scene underscores that the category *woman* has historically been defined primarily by those who wielded institutional and intellectual authority. This insight aligns closely with Simone de Beauvoir's foundational claim that "one is not born, but rather becomes, a woman"–that gender is a socially produced through discourse and representation (1989, 267). It also resonates with Kate Millett's (1969/2000) argument in *Sexual Politics*, which demonstrates how sexual difference is constructed and reinforced through patriarchal ideology embedded in cultural narratives, academic disciplines, and artistic traditions.

Moreover, this passage can serve as a bridge to introducing queer scholarship, particularly Judith Butler's work in *Gender Trouble* (1990) and *Bodies That Matter* (1993), which theorises the concept of gender as being ideologically constructed and performative. Woolf's depiction of the misrepresentation of women by male discourse emphasises the relational nature of discursive power and its constitutive role in shaping both the image and identity of women, whether fictional or real. As such, the text provides a framework for introducing students to feminist concerns in a concrete and reflective manner.

4.1.1. Teaching Strategy: "Canon Audit"

Students can be asked to examine the authors represented in their programme or departmental curriculum and to map patterns of gender representation. This exercise concretises Woolf's critique and highlights the extent to which contemporary curricula continue to reflect inherited structures of epistemic power.

4.2. Deconstructing Binaries and Anticipating the Madwoman Thesis

Woolf's critique of the patriarchal imagination anticipates Sandra Gilbert and Susan Gubar's (1979) argument in *The Madwoman in the Attic* that women in literature are confined to the binary of *angel* or *monster.* Woolf's narrator observes that women in fiction oscillate between "heavenly goodness and hellish depravity" (2004, 96), reflecting not their real lives, but patriarchal fantasy. Her narrator further explains how dramatists limited women to roles defined by their relations to men: "Married against their will, kept in one room, and to one occupation, how could a dramatist give a full or interesting or truthful account of them? Love was the only possible interpreter" (97), and as a result, "The poet was forced to be passionate or bitter" (97). These constraints echo the patriarchal symbolism Gilbert and Gubar (1979) identify, in which women embody either idealised purity or threatening rebellion. This scene also provide a natural entry point to Woolf's short essay "Professions for Women", published in 1931, in which she famously calls on women writers to "kill the Angel in the House" (2017, para. 3)–a metaphor for dismantling these unrealistic yet profound archetypes. In Woolf's (1929, 1931) view, challenging stereotypical representations of women in literature is essential to a woman's literary vocation.

Woolf's narrator also anticipates Gilbert and Gubar's (1979) *madwoman* figure when she speculates that the burning of witches or the persecution of "a wise woman selling herbs" may mark "a lost novelist, a suppressed poet," women driven to madness by thwarted creative energy:

> ...any woman born with a great gift in the sixteenth century would certainly have gone crazed, shot herself, or ended her days in some lonely cottage outside the village, half witch, half wizard, feared and mocked. For it needs little skill in psychology to be sure that a highly gifted girl who had tried to use her gifts for poetry would have been so thwarted and hindered by other people, so tortured and pulled asunder by her own contrary instincts,

> that she must have lost her health and sanity to a certainty. (Woolf 2004, 57)

This passage aligns strikingly with the madwoman thesis: when women's creative energy is blocked by patriarchal repression, it may re-emerge in distorted or destructive forms. Woolf thus provides an interpretive foundation for the feminist analysis of madness, confinement, and silenced creativity. Thus, engaging with Woolf's text allows students to explore and reflect on complex concepts such as the relationship between discourse and power, the discursive construction of gender, and the binary representation of women in male-dominated narratives. This reflection ultimately illuminates the feminist imperative to deconstruct and reconstruct the cultural and literary image of a woman.

4.2.1. Teaching Strategy: Mapping Patriarchal Binaries across Media

To deepen students' understanding of how the angel/ monster binary persists across cultural forms, this activity invites them to analyse representations of women not only in literature, but also in visual art and contemporary advertising. Students can begin by identifying and charting examples of the angel/ monster binary in selected literary texts, noting how female figures are idealised, constrained, or demonised. They then extend this analysis to visual materials–for example, paintings, magazine covers, fashion campaigns, or digital advertisements–tracing how similar archetypes emerge in visual culture. This might include contrasting depictions of women as pure, passive, nurturing, or domestic with portrayals that frame women as dangerous, seductive, unruly, or chaotic.

Once students have identified these patterns, they compare them with Woolf's critique of stereotypical femininity and her call to *kill the angel in the house.* This comparative, cross-media approach encourages students to connect literary analysis with broader feminist concerns, including the discursive construction of gender, the visual policing of women's identities, and the cultural pressures that shape female subjectivity. It also highlights the continued relevance of Woolf's insights by demonstrating how patriarchal binaries persist–and are visually reproduced–in contemporary culture.

4.3. Teaching Gynocriticism through Woolf

Woolf's narrator's attempt to reconstruct a women's literary lineage–through figures such as Lady Winchilsea (1661–1720) (2004, 70), Margaret of Newcastle (1623–73) (71), Dorothy Osborne (1627–95) (73), and Aphra Behn (1640–89) (76)–provides a compelling entry point into the principles of *gynocriticism*. Her reflections on the material and ideological barriers faced by earlier women writers foreground the historical absence of a sustained women's tradition, a lack she identifies as a major impediment to future women authors. This concern anticipates later American feminist critics such as Annette Kolodny (1980), who in "Dancing through the Minefield" insists on the importance of recovering women's writing and revising canons governed by patriarchal values. By assembling an embryonic women's literary heritage, Woolf effectively performs the early work of *gynocriticism* and models its aims: restoration, contextualisation, and recognition of women's shared artistic conditions. Her narrator's catalogue of noblewomen writers underscores Woolf's earlier argument about the necessity of financial independence for artistic production and illustrates the social constraints that enabled or inhibited women's creativity. This gesturing toward canon revision also aligns Woolf with broader feminist interventions in other artistic domains, including the activism of the Guerrilla Girls and feminist art historians who likewise seek to expose the systematic exclusion of women artists.

Woolf's analysis of women's writing further opens a pathway into the methodological concerns of *gynocriticism*. Her assessment of the anger at women's conditions shaping the poetry of Lady Winchilsea and Margaret of Newcastle–diminishing the artistic quality of their work, rendering it "disfigured and deformed by…the same outburst of rage" (Woolf 2004, 71)–alongside her insistence that even "innumerable bad novels" written by women in the eighteenth and nineteenth centuries (75) constitute an essential literary heritage, anticipates arguments later articulated by Elaine Showalter. Showalter's (1981) claim in "Feminist Criticism in the Wilderness" that women's writing expresses a collective cultural experience, an experience that "binds women writers to each other over time and space" (197) resonates strongly with Woolf's assertion that women's books "continue each other, in spite of our habit of judging them separately" (2004,

93) and her reminder that "masterpieces are not single and solitary births" (76). This call for recognising a women's tradition echoes Showalter's (1988/1997) definition of *gynocriticism* in "Towards a Feminist Poetics" as the effort to uncover women's writings to establish a distinct canon and "reconstruct a female framework for the analysis of women's literature" (131).

Woolf's text thus becomes an accessible pedagogical bridge to Showalter's (1977/1999) framework elaborated on in *A Literature of Their Own*, which charts the evolution of women's writing through the Feminine, Feminist, and Female phases. Woolf's observations exemplify each stage of Showalter's framework: the adoption of male pseudonyms (Woolf 2004, 58) and internalised patriarchal values, evident in letters by Dorothy Obsborne (72–73) (Feminine); the rage and protest found in writers such as Lady Winchilsea (68–69) and Charlotte Brontë (81) (Feminist); and the emergence of self-defined artistic autonomy in figures like Jane Austen, Emily Brontë, and Woolf's contemporary, Mary Carmichael (78–79) (Female). As Woolf's narrator notes, Carmichael's "Chloe liked Olivia" (95) encapsulates this final phase, signalling a shift toward centring relationships and experiences between women, independent of male mediation, and anticipating later radical feminist concerns with female bonds and representation.

Through Woolf's *A Room of One's Own*, students gain a clear view of how *gynocriticism* shifts feminist inquiry from economic oppression to textual repression, recovery, and re-evaluation. Analysing Woolf alongside Showalter reveals the coherence of these feminist genealogies and makes visible the ways patriarchal power operates not only through material conditions but also through literary history itself. In this way, Woolf's text not only historicises the evolution of feminist literary criticism but also provides a conceptual and methodological bridge between fictional representation and the lived structures of patriarchal power.

4.3.1. Teaching Strategy: Mini-Archive Task

Students collaboratively build a small digital or physical archive of overlooked women writers or artists, presenting the historical conditions that led to their marginalisation. This exercise enables students to practice feminist methodologies of recovery, canon revision, and critical reclamation–core principles of *gynocriticism*.

5. Language, Subjectivity, and Gendered Expression

A third thematic cluster opened by *A Room of One's Own* concerns the relationship between language, subjectivity, and gendered expression. The narrator's reflections on "a man's sentence" (Woolf 2004, 89) and her observation that some women "wrote as women, not as men write" (87) invite students to consider how linguistic form is shaped by cultural norms, how subjectivity is mediated through available modes of expression, and how women writers have historically struggled to articulate experience within linguistic structures designed by and for men.

5.1. Woolf's "Man's Sentence" and Gendered Language

When Woolf's narrator remarks that the conventional literary sentence "was a man's sentence" (2004, 89), she foregrounds the gendered nature of stylistic conventions and the deep entanglement between language and identity. She observes that traditional prose–"behind it one can see Johnson, Gibbon and the rest"–is shaped by male habits of thought and male experiences of the world, making it "unsuited for a woman's use" (89–90). Women writers, she argues, thus inherit not only a male literary canon but also a male-shaped syntax, form, and genre. This idea provides a powerful starting point for examining gendered stylistic norms, the politics of literary/ artistic form, and the ways signifying systems sustain or constrain subjectivity. Having already explores the implications of discursive power in earlier sections, students can now examine its operation at the level of linguistic form by analysing how Woolf locates this power within the sentence itself.

Woolf's insight that form itself is gendered can be extended beyond literary language to visual culture, particularly film. Here, instructors can introduce Laura Mulvey's (1975) influential theory of the male gaze, discussed in "Visual Pleasure and Narrative Cinema," which argues that classical cinema constructs women as objects of visual pleasure for a presumed heterosexual male spectator. Mulvey contends that cinematic techniques–camera angles, framing, point-of-view shots–produce a *to-be-looked-at-ness* that fixes women within structures of male desire and narrative control. This offers a compelling parallel to Woolf's *man's sentence*: just as prose is shaped by male experience, visual representation is shaped by male spectatorship. Contemporary scholarship on the *female gaze*, such as "Joey Soloway on The Female Gaze" (Soloway 2016) complicates this issues by examining how women filmmakers disrupt, resist, or reconfigure dominant modes of looking. Incorporating Mulvey (1975) and Soloway (2016) into discussions of Woolf's novella enables students to see how gender shapes not only linguistic structures but also visual and aesthetic ones, thereby enriching their cross-disciplinary grasp of feminist theories of representation.

5.1.1. Teaching Strategy: Stylistic & Visual Comparison

This activity encourages students to recognise how gendered structures shape both linguistic and visual forms of representation. Instructors begin by providing short excerpts from Virginia Woolf, Hélène Cixous, and a male author, for instance, Charles Dickens, and ask students to compare syntactic patterns, use of imagery, and narrative perspective. This illuminates the feminist claim that language is neither neutral nor universal. Through this comparative exercise, students observe how literary style is culturally coded and often shaped by gender politics. This foregrounds the feminist insight that language is never neutral: it reflects and reproduces cultural norms, hierarchies, and assumptions about subjectivity.

To expand this textual analysis into a multimodal feminist practice, instructors then introduce a visual comparison. Students watch two short clips: one from a classical Hollywood film and another from a woman filmmaker such as Jane Campion or Sofia Coppola, and then analyse the camera's gaze and point of view, the framing and positioning of women's bodies, and the ways the viewer is invited to identify with certain characters or perspectives. By recognising how patriarchal structures shape both linguistic and visual expression, students develop multimodal feminist literacy. They learn to identify gendered patterns that operate across artistic media and to see how writers and filmmakers challenge, disrupt, or reinvent those forms. This integrative strategy reinforces Woolf's broader claim: that gender is embedded not only in discourse and representation, but in the very forms through which culture expresses meaning.

5.2. Writing the Body: Woolf and Écriture Féminine

Woolf's reflections on gendered language offer an accessible entry point into feminist psychoanalytic theory–particularly Hélène Cixous's (1976) concept of *écriture féminine* as a mode of writing grounded in the body, fluidity, multiplicity, and sensory experience, discussed in her essay "The Laugh of the Medusa‟. Rather than overwhelming students with dense theoretical terminology, instructors can use Woolf's stylistic observations to scaffold the central insights of French psychoanalytic feminism. Woolf's narrator repeatedly notes that women inherit linguistic forms not designed for them: "there was no common sentence ready for her use" (2004, 88). Because the literary tradition has been shaped by male predecessors, the narrator suggests that language, form, and genre themselves bear a masculine imprint. This allows students to grasp, in concrete terms, the psychoanalytic claim that language is culturally and sexually coded rather than neutral or universal.

Within this framework, Cixous's call for women to write their bodies into language becomes particularly tangible as she insists that "woman must write herself…must write about women…must put herself into the text–as into the world and into history–by her own movement" (1976, 875). Woolf anticipates this injunction when her narrator laments on the disappearance of women's gestures, emotions, and daily experiences from cultural memory–"Nothing remains of it all. All has vanished. No biography or history has a word to say about it" (2004, 104). Similarly, Cixous emphasises that women must reclaim their identities through embodied expression: "Almost everything is yet to be written by women about femininity" (1976, 886). Despite their different historical contexts, both writers stress that women must reshape linguistic forms to express experiences long excluded by patriarchal discourse.

The parallels extend to questions of form. Woolf's narrator's urges women to "knock the sentence into shape" (2004, 89) and to cultivate "the habit of writing naturally" (126), prefiguring Cixous's (1976) plea for women to break inherited structures and write through their bodies–producing a language of fluidity, multiplicity, and sensory movement. In Cixous' words, a woman ought to "put herself into the text… by her own movement" (875). Woolf's own narrative style–digressive, rhythmic, and impressionistic–enacts many qualities later associated with *écriture féminine*, enabling students to encounter the theory as embodied literary practice rather than abstract doctrine. Both thinkers also insist on rejecting internalised patriarchal ideals: Cixous exhorts women to "kill the false woman" (880), while Woolf (1931) famously commands in "Professions for Women" that women writers must "kill the Angel in the House" (2017, para. 3).

Cixous's argument that sexuality and textuality must be rejoined–"its infinitive and mobile complexity" (1976, 886)–finds a clear echo in Woolf's hope that "those unrecorded gestures, those unsaid or half-said words" will someday appear in women's writing (2004, 98). Both stress that bodily experience should shape women's linguistic production and encourage women to write with unrestrained freedom: Cixous implores, "let no one hold you back… not even yourself" (877), while Woolf reassures, "So long as you write what you wish to write, that is all that matters" (123).

By drawing out these continuities, instructors can present these complex feminist philosophies not as an isolated or esoteric body of thought but as a meaningful extension of Woolf's project. Students can thus come to understand *écriture féminine* as a practical feminist strategy–grounded in the same structural challenge Woolf identifies: the need to transform inherited discourse in order to articulate women's subjectivities fully.

5.3. Speaking the Body: Woolf and Parler-Femme

Woolf's work can serve as an accessible entry point into Luce Irigaray's (1977/1985) thought, particularly her concept of *parler-femme* (women's speech) as articulated in *This Sex Which Is Not One.* Irigaray conceives *parler-femme* as a form of feminine expression grounded in embodied difference and multiplicity, in contrast to the singular, hierarchical structures of phallocentric discourse. This mode of speech breaks through the constraints of male-dominated linguistic systems, linking women's language intrinsically to the female body. It is characterized by polymorphism, fluidity, and the capacity to express experiences unique to women–qualities that sharply contrast with the unity and rigidity associated with the phallus.

Irigaray emphasises that female sexuality generates distinct expressive possibilities, which she associates with unique creative powers accessible only to women. She writes: “Woman has sex organs more or less everywhere...she finds pleasure almost everywhere...the geography of her pleasure is far more diversified, more multiple in its differences, more complex, more subtle, than commonly imagined” (1985, 28). Sexual difference, for Irigaray, produces a corresponding difference in linguistic potential: for women, language “always in the process of weaving itself, of embracing itself with words, but also of getting rid of words in order not to become fixed, congealed in them” (29). This fluid and generative mode of expression enables women to explore novel and unique forms of creativity.

Woolf anticipates these ideas in her reflections on women’s creative power. Her narrator observes: “But this creative power differs greatly from the creative power of men....It would be a thousand pities if women wrote like men” (2004, 102). Like Irigaray, she emphasises that feminine creativity is distinct not only in content but in form. Woolf’s narrator further suggests that women should write in a distinct manner, “giving things their natural order, as a woman would, if she wrote like a woman” (106). She cites Jane Austen and observes that Austen “devised a perfectly natural, shapely sentence proper for her own use” (89). Like Irigaray, Woolf’s narrator highlights how form must align with the body: “the book has somehow to be adapted to the body” (90).

Woolf’s own narrative style performs the very multiplicity and fluidity that Irigaray associates with *parler-femme*. By engaging with Woolf’s text, students can witness a direct manifestation of feminine creativity, observing how narrative form and style reflect embodied difference. This loop between theory and literary experience provides a concrete means of understanding Irigaray’s arguments about the connection between female bodies, sexual difference, and distinct modes of expression. In this way, Woolf’s writing becomes both a pedagogical tool and a model for *parler-femme*, grounding abstract feminist theory in a tangible reading experience.

5.4. Teaching through Form: An Embodied Reading Practice

Woolf’s narrative experimentation offers an ideal platform for exploring the formal dimensions of feminist writing. Students can analyse how techniques such as fragmentation, stream of consciousness, shifts between fact and fiction, and impressionistic detail enact what later feminist theorists identify as features of *écriture féminine*. These strategies not only challenge conventional narrative structures but also model alternative modes of subjectivity and expression, highlighting the intimate link between literary form and embodied experience.

By the time students encounter psychoanalytic feminist theory, Woolf has already prepared them to ask its central questions: *What kinds of subjectivities are possible within inherited linguistic forms? How can women reshape literary conventions to articulate embodied experience? What does it mean to write outside patriarchal discourse?* Through close reading of Woolf's text, students can trace how narrative experimentation enacts answers to these questions, revealing the inseparability of style, form, and feminist inquiry.

5.4.1. Teaching Strategy: Embodied Writing Exercise

Students may be asked to rewrite a brief passage from a canonical male author in a *woman's language*, experimenting with fluidity, interiority, digression, or rhythm. By engaging directly with linguistic form, students may experience the conceptual shift Woolf describes, learning firsthand the ways of how form can embody feminist insight. This exercise fosters multimodal literacy, linking textual analysis, embodied practice, and theoretical reflection, and enables students to integrate the principles of feminist thought into their own writing and interpretive strategies. Through this embodied approach, Woolf's text functions not only as a source of ideas but as a model for the lived practice of feminist theory, showing that critical understanding emerges not only through intellectual engagement but also through the imaginative enactment of alternative forms of expression.

6. Conclusion

This article has argued that *A Room of One's Own* offers a uniquely effective pedagogical bridge between literature and feminist theory, enabling students to approach complex philosophical arguments through the embodied, rhetorical, and narrative strategies of Virginia Woolf's text. By situating feminist concerns–material inequality, gendered subjectivity, discursive power, and linguistic embodiment–within a vivid literary framework, Woolf anticipates key strands of feminist thought while providing a concrete context through which students can grasp them.

Close reading of Woolf's reflections on language, creativity, and gendered experience equips students with conceptual tools for analysing the intersections of gender, class, language, discourse, and power. Woolf's narrative thus becomes an accessible entry point into diverse trajectories of feminist theory, ranging from materialist and socio-political critiques to psychoanalytic and poststructural approaches such as *écriture féminine* and *parler-femme*. The parallels between Woolf's stylistic choices and the theories later articulated by Cixous (1976) and Irigaray (1977) demonstrate how literary form can illuminate abstract concepts–fluidity, multiplicity, embodied expression–in ways that are immediate and experientially grounded. Through Woolf's digressive, rhythmic, and imaginative prose, students encounter an enacted version of these theories, reinforcing conceptual understanding through literary experience.

More broadly, integrating Woolf into the teaching of feminist theory underscores the pedagogical value of interdisciplinary learning. Literature enables students to encounter theoretical arguments not as detached abstractions but as lived and narrated concerns, fostering critical reading skills, analytical nuance, and reflective awareness of how personal experience and structural inequality shape the production of knowledge–a central tenet of feminist epistemology. Engaging with fiction also makes theoretical material more accessible and memorable, encourages interdisciplinary connections across philosophy, psychology, cultural studies, and history, and invites creative and imaginative engagement with questions of gender and power.

By tracing feminist traditions in Woolf's text, this article has sought to demonstrate how literary analysis can enrich the study of feminist theory and deepen students' understanding of how language, gender, and power operate in both historical and contemporary contexts. In doing so, it contributes to broader conversations in feminist pedagogy about how best to cultivate meaningful, interdisciplinary, and transformative learning experiences. Ultimately, teaching feminist theory through literature does more than illuminate feminist histories: it equips students with the intellectual flexibility, critical sensitivity, and empathetic imagination needed to engage thoughtfully with gendered realities in the world today.

References

Beauvoir, Simone de. 1989. *The Second Sex*, translated by H. M. Parshley. New York: Vintage Books, first published 1949.

Butler, Judith. 1993. *Bodies That Matter: On the Discursive Limits of Sex.* New York: Routledge.

Butler, Judith. 1990. *Gender Trouble: Feminism and the Subversion of Identity.* New York: Routledge.

Cixous, Hélène. 1976. "The Laugh of the Medusa," translated by K. Cohen, P. Cohen. *Signs*, 1(4): 875–93.

Coles, Robert. 1989. *The Call of Stories: Teaching and the Moral Imagination.* Boston: Houghton Mifflin.

Combahee River Collective. 1981. A Black Feminist Statement. In *This Bridge Called My Back: Writings by Radical Women of Color*, edited by C. Moraga and G. Anzaldua, 210–18. Watertown, Massachsuetts: Persephone Press (first published 1977).

Cunningham, Catriona, and Jennie Mills. 2024. "Glow up: The Power of Fiction in Higher Education Research." *Teaching in Higher Education*, 29(7): 1879–96. https://doi.org/10.1080/13562517.2024.2359700

Eagleton, Marry. 1991. *Feminist Literary Criticism.* London: Routledge. https://doi.org/10.4324/9781315846163

Egan, Kieran. 1997. *The Educated Mind: How Cognitive Tools Shape Our Understanding.* Chicago: University of Chicago Press.

Federici, Silvia. 2004. *Caliban and the Witch: Women, the Body and Primitive Accumulation.* New York: Autonomedia.

Freire, Paulo. 1986. *Pedagogy of the Oppressed.* New York: Continuum.

Gilbert, Sandra M., and Susan Gubar. 1979. *The Madwoman in the Attic: The Woman Writer and the Nineteenth-Century Literary Imagination.* New Haven, Connecticut: Yale University Press.

Hansen, Thomas Illum. 2023. "Phenomenological Exploration in Literature Education: On the Theoretical Development of a Phenomenological Approach to Inquiry-based Literature Teaching as a Focal Point for a Large-scale Intervention Study in Denmark." *LI-Educational Studies in Language and Literature*, 23(1): 1–26. https://doi.org/10.21248/l1esll.2023.23.1.382

Haraway, Donna. 1988. “Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective.” *Feminist Studies*, 14(3): 575–99. https://doi.org/10.2307/3178066

Harding, Sandra. 1987. Introduction: Is There a Feminist Method? In *Feminism and Methodology: Social Science Issues,* edited by S. Harding, 11–14, Bloomington, Indiana: Indiana University Press.

Hemmings, Clare. 2011. *Why Stories Matter: The Political Grammar of Feminist Theory.* Durham and London: Duke University Press. https://doi.org/10.2307/j.ctv1220mp6

Hesse-Biber, Sharlene N. 2012. *Handbook of Feminist Research: Theory and Praxis*, 2nd ed., Thousand Oaks, California: SAGE Publications, Inc. https://doi.org/10.4135/9781483384740

Holmes, Diana. 2012. “Feminist Literary Theory and the Woman Reader.” *Tropelías: Revista de Teoría de la Literatura y Literatura Comparada*, 36: 51–68. https://doi.org/10.26754/ojs_tropelias/tropelias.20213653199

Hooks, Bell. 1982. *Ain’t I a Woman: Black Women and Feminism.* Boston: South End Press.

Hooks, Bell. 1984. *Feminist Theory: From Margin to Center.* Boston: South End Press.

Irigaray, Luce. 1985. *This Sex which Is Not One*, translated by C. Porter, C. Burke. Ithaca, New York: Cornell University Press. first published 1977.

Jaggar, Alison. 1983. *Feminist Politics and Human Nature.* Brighton: Harvester Wheatsheaf.

Kolodny, Anette. 1980. “Dancing through the Minefield: Some Observations on the Theory, Practice, and Politics of a Feminist Literary Criticism.” *Feminist Studies*, 6(1): 1–25.

Kier-Byfield, Sophia. 2024. Close Reading: Critical Feminist Method and Pedagogical Process. In *Intersectional Feminist Research Methodologies: Applications in the Social Sciences and Humanities*, edited by Jennifer Cooke and Line Nyhagen. 176–88. London: Routledge.

Lorde, Audre. 1984. *Sister Outsider: Essays and Speeches.* Freedom, California: Crossing Press.

Lukić, Jasmina, and Adelina Sánchez-Espinosa. 2019. Feminist Perspectives on Close Reading. In *The Routledge Companion to Feminist Literary Criticism*, edited by Mary Eagleton and Emma Parker, 123–35. London: Routledge.

Lykke, Nina. 2010. *Feminist Studies: A Guide to Intersectional Theory, Methodology and Writing.* New York: Routledge.

Marek, Kate. 2006. “Using Literature to Teach in LIS Education: A Very Good Idea.” *Journal of Education for Library and Information Science*, 47(2): 144-59. https://doi.org/10.2307/40324328

Millett, Kate. 2000. *Sexual Politics.* Urbana, Illinois: University of Illinois Press. first published 1969.

Moi, Toril. 1985. *Sexual/Textual Politics: Feminist Literary Theory.* London: Methuen.

Mulvey, Laura. 1975. “Visual Pleasure and Narrative Cinema.” *Screen*, 16(3): 6–18. https://doi.org/10.1093/screen/16.3.6

Nochlin, Linda. 1971. “Why Have There Been No Great Women Artists?” *Art News,* 69(9): 22–39, 67–71 (reprinted in *Women, Art, and Power and Other Essays*, 145–78. New York: Harper & Row, 1988).

Olsen, Tillie. 1978. *Silences.* New York: Delacorte Press.

Rich, Adrienne. 1971. “When We Dead Awaken: Writing as Re-Vision.” *College English*, 34(1): 18–30.

Rosenblatt, Louise M. 1994. *The Reader, the Text, the Poem: The Transactional Theory of the Literary Work.* Carbondale, Illinois: Southern Illinois University Press.

Showalter, Elaine. 1981. “Feminist Criticism in the Wilderness.” *Critical Inquiry*, 8(2): 179–205.

Showalter, Elaine. 1985. *The New Feminist Criticism: Essays on Women, Literature, and Theory.* New York: Pantheon Books.

Showalter, Elaine. 1997. Towards a Feminist Poetics. In *Twentieth-Century Literary Theory, e*dited by K. M. Newton. 216–20. London: Macmillan Publishers Limited (first published 1988).

Showalter, Elaine. 1999. *A Literature of Their Own: British Women Novelists from Bronte to Lessing.* Princeton New jersey: Princeton University Press (first published 1977).

Sidney, Philip. 1973. *An Apology for Poetry (Or The Defence of Poesy)*, edited by G. Shepherd and R. W. Maslen. Manchester: Manchester University Press (first published 1595).

Soloway, Joey. 2016. “Joey Soloway on The Female Gaze.” *Master Class TIFF.* https://www.youtube.com/watch?v=pnBvppooD9I/

Williams, Kevin, and Patrick A Williams. 2023. “The Rationale for the Teaching of Literature: Soundings in Paul Hirst’s Epistemology.” *Journal of Philosophy of Education*, 57(1): 276–92. https://doi.org/10.1093/jopedu/qhad009

Wollstonecraft, Mary. 1999. *A Vindication of the Rights of Women.* Boston: Bartleby (first published 1792).

Woolf, Virginia. 1992. *Women & Fiction: The Manuscript Version of “A Room of One’s Own*, edited by S. P. Rosenbaum. Oxford: Shakespeare Head Press by Blackwell Publishers.

Woolf, Virginia. 2004. *A Room of One’s Own.* London: Penguin Books.

Woolf, Virginia. 2017. “Professions for Women.” *National Society for Women’s Service.* http://s.spachman.tripod.com/Woolf/professions.htm.

Yazell, Bryan., Petersen, Klaus., Marx, Paul. et al. 2021. “The Role of Literary Fiction in Facilitating Social Science Research.” *Humanit Soc Sci Commun*, 8:261. https://doi.org/10.1057/s41599-021-00939-y

Organon F 33 (1) 2026: 106–118 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2026.33104

RESEARCH ARTICLE

Bertrand Russell and the Ethics of Truth: A Critique of Pragmatism from James to Orwell

Evgeni Latinov*

Received: 8 June 2025 / Accepted: August 11, 2025

Abstract: This article examines Bertrand Russell's engagement with pragmatism through his critical yet respectful treatment of William James and John Dewey in *A History of Western Philosophy* (1945). While Russell endorses the metaphysical innovation of neutral monism, especially in James's doctrine of radical empiricism, he rejects the pragmatist redefinition of truth as a function of practical success. The analysis reconstructs Russell's main arguments against pragmatist epistemology, emphasizing his insistence on truth as correspondence between belief and reality. Particular attention is given to Russell's theory of belief, his conception of the "bearer of truth," and his critique of instrumentalist approaches that conflate usefulness with validity. The article concludes by drawing a parallel with George Orwell's *Nineteen Eighty-Four*, showing how Russell's concerns anticipate the political and ethical risks of subordinating truth to utility. In light of current challenges such as misinformation, algorithmic manipulation, and post-truth politics, Russell's critique emerges as a timely defense of objective truth and the conditions for free inquiry and public discourse.

Keywords: Bertrand Russell; John Dewey; objectivity; pragmatism; theory of truth; William James.

* Sofia University St. Kliment Ohridski
https://orcid.org/0000-0003-2358-1701
Faculty of Philosophy, Department of Logic, Ethics and Aesthetics, 15, Tsar Osvoboditel Blvd, Sofia 1504, Bulgaria
elatinov@phls.uni-sofia.bg

1. Introduction

The early decades of the twentieth century marked a turning point in the development of modern philosophy. Among the intellectual movements that emerged in response to the perceived limitations of classical metaphysics and absolute idealism, pragmatism stood out for its radical redefinition of traditional concepts such as meaning, knowledge, and especially truth. Pragmatists like William James and John Dewey sought to ground philosophy not in abstract speculation, but in experience, action, and consequence. They proposed that beliefs should be treated as tools for coping with life–instruments for navigating reality rather than static representations of it. For both thinkers, the truth of a belief was to be measured not by its correspondence to external facts, but by its capacity to function effectively in guiding action.

Bertrand Russell, a central figure in the rise of analytic philosophy, responded to pragmatism with a combination of admiration and strong criticism. In the penultimate and preceding chapters of his *History of Western Philosophy* (1945), devoted respectively to William James and John Dewey, Russell presents a sustained polemic against the pragmatist theory of truth–which he considers its most original and most problematic doctrine. At the same time, he expresses a deeply respectful and even sympathetic attitude toward both philosophers. He acknowledges their originality, moral seriousness, and their shared dissatisfaction with the speculative excesses of idealism. Most notably, he praises one key pragmatist idea with which he substantially agrees: the doctrine of *neutral monism*.

Russell's critique of pragmatism must be situated within the broader intellectual context of the late nineteenth and early twentieth centuries. While he shared the pragmatist impulse to reject metaphysical absolutism and engage with lived experience, he was equally committed to preserving a conception of truth grounded in objective correspondence with reality. This commitment to realism would come to define much of analytic philosophy in the decades that followed. The two chapters on James and Dewey thus offer a valuable perspective on Russell's complex relationship to the pragmatist tradition–a relationship shaped by both philosophical convergence and principled disagreement.

This article explores that relationship by focusing on the two central ideas that structure Russell's engagement with pragmatism: his rejection of the pragmatist redefinition of truth, and his partial endorsement of neutral monism. The discussion reconstructs Russell's arguments against the utility theory of truth, explores his metaphysical agreement with James's radical empiricism, and highlights his evolving account of belief and the "bearer of truth." In conclusion, the paper extends Russell's warnings about the political implications of redefining truth, drawing a parallel with Orwell's dystopian critique of ideological language, and briefly reflects on contemporary challenges posed by misinformation, algorithmic discourse, and AI-generated realities.

2. William James and the Doctrine of Neutral Monism

According to Bertrand Russell, the three principal figures associated with pragmatism are William James, John Dewey, and F. C. S. Schiller. Among these, Russell considers Schiller of lesser philosophical significance and focuses primarily on James and Dewey. Notably, he does not include Charles Sanders Peirce in this list, despite Peirce's foundational role in articulating the pragmatist maxim. In *A History of Western Philosophy*, Peirce appears only indirectly: first, when James is credited with adopting Peirce's principle that the meaning of an idea is to be clarified by its practical consequences (Peirce 1878); and second, in Russell's mention that Dewey approvingly cites Peirce's definition of truth as "that opinion which is fated to be ultimately agreed to by all who investigate."

James's prominence in Russell's discussion is not incidental. Beyond his historical significance, James is singled out for his formulation of the doctrine he initially termed *radical empiricism*, and which his followers later came to call *neutral monism*. It is this metaphysical innovation that Russell finds not only original but deeply compelling.

At the heart of neutral monism lies a rejection of the classical dichotomy between subject and object–a dichotomy central to Western philosophy from Socrates through Descartes. The traditional epistemological model assumes that knowledge is a relation between a subject (the knower) and an object (the known) and is closely tied to ontological dualism: the subject is aligned with mind, the object with matter. Neutral monism challenges the metaphysical primacy of this distinction. It posits instead that reality is composed of a single type of fundamental "stuff"–what James called *pure experience*–which can function either as subject or object depending on relational context.

James's radical move is to deny that consciousness is a metaphysically independent entity. Rather than being a substance distinct from the world, consciousness is a functional property of certain configurations of pure experience. A given portion of experience can serve as the perceiver in one context and the perceived in another. There is no inherent ontological divide between mental and physical, subjective and objective; such distinctions emerge from relational patterns rather than from fundamental categories.

Russell not only embraces the core insight of neutral monism but describes his shift toward it as a philosophical conversion. In a pivotal passage, he writes: "For my part, I am convinced that James was right on this matter, and would, on this ground alone, deserve a high place among philosophers" (Russell 1945).

Despite this endorsement, Russell raises a terminological concern. The term *experience*, he notes, is too closely associated with human psychology and may inadvertently suggest a form of idealism. For common sense, experience is something living beings have–yet for Russell, the fundamental constituents of the world must be applicable to both animate and inanimate reality. He therefore prefers more neutral terms, such as *events* or *occurrences*.

Nonetheless, the philosophical significance of neutral monism remains. By abandoning the rigid opposition between mind and matter, neutral monism provides a metaphysical foundation compatible with both scientific realism and a naturalistic account of consciousness. It offers a third alternative to the traditional dichotomy between idealism and materialism. Russell's own later work–including his theory of perception, his analysis of sense-data, and his causal theory of knowledge–bears the mark of this influence. Although he would formulate neutral monism more rigorously than James and adopt different terminology, the metaphysical core of the doctrine remains intact.

Thus, while Russell would later turn to critique James's theory of truth, he remains profoundly indebted to James's insight into the structure of reality. Neutral monism stands as one of the few areas where Russell finds deep and lasting agreement with the pragmatist tradition.

3. Russell's Critique of James's Theory of Truth

While Bertrand Russell expresses admiration for William James's metaphysical insights, especially the doctrine of neutral monism, he unequivocally rejects the pragmatist redefinition of truth–the idea that truth is what "works." According to the core pragmatist thesis, a belief is true if it proves practically effective: if it leads to satisfactory, beneficial, or otherwise desirable outcomes in experience. This is the defining feature of pragmatism, shared by both James and Dewey, and the source of its name. For Russell, however, this theory undermines the very foundations of epistemology.

In *Pragmatism* (1907), James writes: "The true is only the expedient in the way of our thinking, just as the right is only the expedient in the way of our behaving" (James 1907, 28). He later adds that if belief in God "works satisfactorily in the widest sense of the word, then it is true." In other words, beliefs are true not because they correspond to facts, but because they help us live better, solve problems, or feel fulfilled. Russell considers this claim not only mistaken but dangerous.

He develops three main arguments against it:

> **Epistemic Unmanageability.** Russell first argues that the pragmatist theory is unworkable in practice. To determine whether a belief is true, one must assess the consequences of believing it and evaluate whether those consequences are beneficial. For example, consider the proposition: "Columbus crossed the Atlantic in 1492." According to the pragmatist, we must judge whether believing this yields good outcomes. But this is absurd. One might get a higher grade on an exam for giving the right date–a good outcome. Yet for someone else, this might mean losing a scholarship due to lower relative performance–a bad outcome. Consequences differ across contexts and individuals. Truth becomes interest-relative and subjective, no longer grounded in stable criteria.
>
> **Infinite Regress.** Second, Russell shows that pragmatism leads to a vicious regress. If a belief is true because it has good consequences, then

we must evaluate whether those consequences themselves are good. But to do that, we must form additional beliefs–beliefs about the consequences of the original belie–which must also be evaluated. This process never ends. Each layer of belief requires another to justify it, leading to an infinite regress that makes epistemic grounding impossible.

Ontological Confusion. Most damningly, Russell argues that pragmatism severs truth from reality. If truth is determined by usefulness, then beliefs about non-existent entities could be deemed true. For instance, belief in Santa Claus may produce joy in children, but that does not make "Santa Claus exists" a true proposition. Truth, Russell insists, must be tied to what causes a belief–namely, the actual existence of the entity in question–not to the effects it produces. The belief that Columbus existed is true because Columbus actually lived, not because believing in him brings satisfaction.

Russell sees James's theory as driven by a desire to vindicate religious faith. By equating truth with practical success, James creates a conceptual space in which belief in God can be validated by its psychological benefits. But this reverses the logical order. A believer does not say, "I believe in God because it comforts me." Rather, they say, "I am comforted because I believe that God exists." The belief is held to be true first, and only then does emotional comfort follow. To invert this relationship, Russell argues, is to confuse epistemology with psychology.

Moreover, Russell contends that if truth is what serves human needs, then truth becomes susceptible to ideological manipulation. Those in power–political, religious, or corporate–can define what is "useful" and, therefore, what is "true." Pragmatism, thus construed, opens the door to propaganda and relativism. Russell's concern is not merely theoretical; he sees this conception of truth as a threat to intellectual freedom and scientific integrity.

Despite his admiration for James's ingenuity and courage, Russell concludes that the pragmatist theory of truth must be rejected. Truth must be objective, grounded in a belief's correspondence to facts–not in its psychological or pragmatic utility. Otherwise, philosophy forfeits its critical function and becomes an instrument of persuasion or comfort.

4. John Dewey

The two central principles of pragmatism discussed in relation to William James–he doctrine of neutral monism and the redefinition of truth–also appear in the philosophy of John Dewey, though in distinct and more systematized forms. While Dewey does not explicitly formulate a metaphysical doctrine akin to James's "pure experience," his conception of the organism and its relation to the environment implicitly rests on a metaphysical outlook that closely resembles neutral monism. As with James, Bertrand Russell accepts this metaphysical core but forcefully critiques the associated theory of truth.

Dewey's philosophy centers on the concept of *inquiry*, which he frames in biological and functional terms. As Russell interprets him, Dewey begins with the idea of an organism–not merely a physical body or a detached mind, but a unified, adaptive entity. When an organism interacts with its environment, the result may be satisfactory or unsatisfactory. In cases of failure or disruption, adjustment must occur. If the environment is altered, the process is external. But if the organism changes internally–by adjusting its beliefs, habits, or understanding–this process is what Dewey calls *inquiry*. Inquiry is thus an adaptive mechanism, a way for the organism to reestablish equilibrium through internal reorganization. Russell summarizes this view with a vivid military analogy:

> During a battle, your primary concern is to change the environment – namely, to defeat the enemy. But during reconnaissance before the battle, your main concern is to adjust your own forces in light of the enemy's position. This earlier phase is the phase of "inquiry." (Russell 1945)

This conception of inquiry, grounded in evolutionary biology and behavioral psychology, leads directly to Dewey's *instrumentalist theory of truth.* Beliefs, in this framework, are not representations of an independent reality but tools for successful adaptation. Truth becomes a function of problem-solving: a belief is true to the extent that it contributes to resolving the disequilibrium that prompted inquiry. In contrast to the traditional correspondence theory, Dewey's account is dynamic, contingent, and contextual.

Russell finds much to admire in Dewey's underlying metaphysics. He sees Dewey's notion of the organism as implicitly neutral monist: mental and physical phenomena are not distinct substances but different aspects of one unified system. Beliefs, in this picture, are not immaterial ideas floating in a separate mental realm but dispositions to act–embedded in behavior and shaped by environmental interaction.

To illustrate this, Russell offers two striking examples. First, imagine someone at a zoo who hears over the loudspeaker that a lion has escaped. Even without seeing the lion, the person reacts with fear and heightened alertness. This reaction–physiological and behavioral–constitutes the belief that a lion has escaped. Whether based on direct perception or linguistic input, the organism's state expresses the same belief. Second, consider someone descending a staircase in the dark. They do not consciously think, "There is a step beneath my foot," but step forward confidently. The belief in the continued presence of steps is embedded in their action. Only when the steps end unexpectedly and they stumble does the belief become explicit. This illustrates that beliefs can exist unconsciously and non-verbally, as tendencies or orientations within the organism.

Up to this point, Russell and Dewey are largely in agreement. But the divergence appears when we ask: *what makes a belief true*?

For Russell, a belief is true if it corresponds to a fact–if the state of the world matches the propositional content of the belief. Returning to the battlefield analogy: suppose reconnaissance reports that the enemy is concentrating forces on the left flank. If the general acts on this belief and wins the battle, that success alone does not make the belief true. The belief is true if and only if the enemy was indeed massing forces on that flank–regardless of the outcome.

Dewey, by contrast, evaluates truth based on success in inquiry. A belief is validated not by its match to an independent reality, but by its utility in guiding action and resolving problems. If the general wins the battle, then–from Dewey's perspective–the belief may be said to have "worked," and thus is justified. This shift from correspondence to function is where Russell mounts his strongest critique.

He argues, as he did against James, that *success is an unreliable guide to truth.* A belief might lead to success by accident, or fail due to unrelated factors. In the staircase example, a person might descend safely while holding a false belief, or fall despite having a true one. Truth, for Russell, cannot be equated with practical effectiveness.

Moreover, Russell warns that Dewey's instrumentalism risks *relativizing historical truth.* If truth is what works, then the truth of a historical claim depends on its present-day utility, not on what actually happened. This opens the door to ideological manipulation: narratives about the past may be judged "true" based on their political or educational effectiveness, rather than their fidelity to fact. For Russell, this is epistemologically and ethically unacceptable. Historical propositions–such as "Caesar crossed the Rubicon in 49 BCE"–must be assessed according to whether they correspond to past events, not whether they serve current goals.

This concern leads to a deeper philosophical question: *what is the proper bearer of truth?* Russell insists that it is not sentences, utterances, or even statements in themselves, but *beliefs*–mental states with propositional content, held by conscious agents. Words may express a belief, but only when they are spoken with conviction and understanding. Repeating a sentence without grasping its meaning–as in the case of parroting "A lion has escaped from the zoo" without awareness–does not constitute a belief, and thus cannot be evaluated for truth or falsity.

Russell's emphasis on the *bearer of truth* reflects his commitment to a realism grounded in the structure of belief. Even though he finds Dewey's metaphysical framing compatible with his own–particularly in its rejection of Cartesian dualism–he rejects Dewey's attempt to ground truth in the fluid, instrumental outcomes of inquiry.

For Russell, *truth must remain correspondence with fact.* It is not a flexible tool to be judged by what "works," but a relation between thought and reality. While inquiry may be the method through which we arrive at true beliefs, it does not itself define truth. In abandoning this principle, Russell believes Dewey risks turning philosophy into a servant of expediency.

5. The Political Stakes of Truth: Russell, Orwell, and the Ethics of Objectivity

Bertrand Russell's critique of pragmatism culminates in a warning that extends far beyond epistemology. While his disagreements with William James and John Dewey focus on the redefinition of truth as a matter of practical success, Russell ultimately identifies deeper ethical and political dangers. If truth is untethered from correspondence with reality and is instead defined by what "works" or what is useful, it becomes vulnerable to manipulation – particularly by those in positions of power. For Russell, this represents not merely a philosophical error but a civilizational threat.

In the final sentences of *A History of Western Philosophy*, Russell reflects on the effects of modern scientific and technological advancement. He warns: "Modern technique has revived the sense of collective power among human communities...I feel here a grave danger–the danger of what I might call cosmic pride" (Russell 1945).

This *cosmic pride*–the illusion that humanity can not only act upon the world but reshape its moral and epistemic foundations–tempts us to treat truth as a tool of will rather than a constraint upon it. If truth is redefined as what serves our purposes, then those with the most power are best positioned to define what counts as true.

Russell fears that pragmatism's instrumentalist conception of truth encourages precisely this illusion. In divorcing truth from a mind-independent reality, pragmatism offers a seductive but ultimately corrosive vision of human beings not as discoverers of truth, but as its creators. The consequences of this view are powerfully dramatized in George Orwell's dystopian novel *Nineteen Eighty-Four* (1949), published just four years after Russell's work.

In Orwell's imagined regime, the ruling Party rewrites history and language in order to control perception and belief. The Party's chilling slogan makes the stakes explicit: "Who controls the past controls the future: who controls the present controls the past" (Orwell 1949).

In this world, facts are fluid, and truth is reduced to political expediency. There are no stable realities–only shifting narratives dictated by authority. Language, belief, and memory no longer aim to reflect the world but to reinforce domination. Russell, though writing as a philosopher rather than a novelist, anticipates this epistemological totalitarianism. He sees in pragmatism's reduction of truth to practical success the philosophical foundation for Orwell's nightmare.

To be clear, Russell does not accuse James or Dewey of authoritarian intent. On the contrary, he regards them as morally earnest thinkers. But he worries that the logic of pragmatism, taken to its conclusion, erodes the very idea of truth as something objective and independent of human aims. Once truth is defined by utility, it becomes pliable. And once it becomes pliable, it can be engineered–not only by governments, but by corporations, media systems, and digital algorithms.

This concern is no longer hypothetical. In the contemporary world, the distinction between truth and falsehood has become increasingly fragile. Social media platforms privilege engagement over accuracy. Algorithmically curated content creates epistemic bubbles. Misinformation spreads at unprecedented speed, often generated and amplified by artificial intelligence. Deepfakes blur the line between reality and fabrication. In this context, Orwell's post-truth dystopia has moved from fiction to daily experience.

Russell's defense of truth as correspondence–as fidelity to what is–stands as a moral imperative. Without this anchor, public discourse becomes hollow, scientific inquiry collapses into consensus-building, and the possibility of rational disagreement vanishes. Even in democratic societies, when truth is subordinated to utility–in advertising, politics, or journalism–the result is not pluralism, but manipulation. What matters is no longer *what is true*, but *what works*–for branding, for ideology, for influence.

Russell's critique of pragmatism, then, is not merely a disagreement about epistemological theory. It is a call to intellectual and civic responsibility. Truth must not be equated with what is persuasive, beneficial, or emotionally satisfying. It must remain accountable to a world that exists independently of our interests and desires. Only then can reasoned communication, free inquiry, and ethical integrity be preserved.

In defending objectivity, Russell aligns not only with traditional philosophical realism but with the ethical commitments of a democratic society. As Orwell understood, and as Russell foresaw, the collapse of truth into utility endangers both freedom and thought. The stakes of this philosophical debate are nothing less than the preservation of truth itself as a cultural and moral ideal.

6. Conclusion

Bertrand Russell's engagement with pragmatism reveals a principled defense of the conditions under which knowledge and public discourse remain possible. While he appreciated William James's neutral monism and John Dewey's biologically grounded inquiry, Russell rejected the pragmatist redefinition of truth as merely what is useful to believe. For him, truth must retain its objective status as a relation between belief and fact, not a function of practical success.

His concern extended beyond theory to ethical and political dangers that arise when truth is reduced to utility–dangers dramatized in Orwell's vision of totalitarian control over language and memory. In today's era of algorithmic misinformation, politicized narratives, and automated content generation, Russell's warnings remain urgently relevant.

Defending truth as correspondence is not nostalgic realism but a foundational principle of free thought, democratic culture, and philosophical integrity. Without this commitment, the intellectual and ethical bases of science, philosophy, and democracy risk erosion. Russell's critique serves as a vital reminder that truth must transcend utility to preserve both accuracy and freedom.

Funding

This study was financed by the European Union – NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No. BG-RRP-2.004-0008-C01.

References

James, William. 1907. *Pragmatism: A New Name for Some Old Ways of Thinking.* New York: Longmans, Green, and Co.

Orwell, George. 1949. *Nineteen Eighty-Four.* London: Secker & Warburg.

Peirce, Charles Sanders. 1878. “How to Make Our Ideas Clear.” *Popular Science Monthly*, 12 (January): 286–302.

Russell, Bertrand. 1945. *A History of Western Philosophy*. London: George Allen & Unwin.

Organon F 33 (1) 2026: 119–142 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2026.33105

RESEARCH ARTICLE

Vagueness: Two Myths

Jeffrey J. Watson*

Received: 7 August 2025 / Revised: 2 December 2025 / Accepted: 8 December 2025

Abstract: Epistemicism about vagueness is the position that bivalence holds for every instance of a vague predicate, even if truth or falsity is unknowable in borderline cases. Epistemicism is accused of rejecting the *tolerance intuition*, and committing itself to *sharp borderlines*. Mainstream Epistemicists, like Williamson and Sorensen, accept these accusations as costs of their view. I argue instead that both are myths. First, I argue our intuitions support only *generic, dense* tolerance principles, which are non-paradoxical. Epistemicists can affirm these principles, without inferring any paradoxical principle, and so can embrace the tolerance intuition. Second, bivalence is perfectly compatible with the denial of sharp borderlines, provided that we model the extension of vague predicates as scattered stochastically and non-monotonically across a gradient, just as we should expect if meaning depends on use. My revisionary form of epistemicism better balances our intuitions about vagueness with the conservation of bivalence.

Keywords: borderline cases; continuous sorites; epistemicism; sorites; tolerance principle; vagueness.

* Arizona State University
https://orcid.org/0000-0003-0412-8653
Arizona State University, School of Historical, Philosophical, and Religious Studies, 975 S Myrtle Ave., Mail Code 4302, Tempe, AZ 85287-4302, USA
jjwatso2@asu.edu

1. Introduction

Vague predicates give rise to the sorites paradox. The conventional formulation of the sorites paradox arrives at its conclusion either by a large but countable series of *modus ponens* steps, or by mathematical induction. We begin by quantifying over a domain D which is well-ordered by ≤, like ℕ, the set of natural numbers. Any monotonic vague relation which can be represented by ≤ will do, such as "as short or shorter than," "of an equal or less intense shade of yellow than," or "as rich or less rich than." For instance, our domain might be the set of possible heads, ordered by numbers of hairs on a head. We then offer an instance of the following:

(Well-ordered Sorites) *For predicate F, base case b, and some undetectably small* $c > 0$*, where* $c \in C$ *and* $|C| \leq |\mathbb{N}|$*:*

(P1) Fb
(P2) $(\forall x)(\forall y, |y| \leq c)(\ Fx \ \rightarrow F(x+y))$
(C) $(\forall x)(Fx)$

For example, using "number" to mean "natural number," Fx = "Somebody with x number of hairs is bald," and letting b=0 and c=1:

(P1) Someone with zero hairs is bald
(P2) For any number of hairs, if someone with that number of hairs is bald, then someone with 1 more hair is also bald.
(C) Anyone with any number of hairs is bald. (Even someone with 10 trillion hairs).

We can make the domain as fine-grained as we please, such as including 24.99% of a hair and 25% of a hair, or any other difference so small as to be phenomenologically undetectable through perception or imagination. We can't, however, include all of the rational or real number values *between* 24.99% and 25%, since the rationals and reals are not well-ordered by ≤. In the conventional paradox, the domain must consist of well-ordered units of measure, since otherwise the use of mathematical induction or a finite modus ponens series would be invalid.

There are many proposals for avoiding this paradox, usually by motivating some rejection of the *tolerance principle* (P2), despite its appearing intuitive. First, it's widely agreed that vague predicates are sensitive to conversational context and the purposes of a conversation: what's fast for a marathon is not fast for a fighter jet. So, *contextualists* argue that there are micro-shifts in the conversational context at each step of the sorites, on each application of "fast" (Raffman 1994; 2005a; 2005b), making (P2) equivocal. While I accept that vague predicates are context-sensitive, it seems to me that even if context were kept rigidly fixed for an idealized jury of competent speakers, the jury would struggle to come up with a consensus on $\mathrm{F}n$ & $\sim\mathrm{F}n_{+1}$ for any *n*. So, for the purpose of this paper, I assume contextualism alone is insufficient to resolve the paradox (Soames 1998; Stanley 2003; Keefe 2007).

Second, the tolerance principle might be rejected by *denying bivalence.* For instance, we might reject (P2) by holding that truth conditions are indeterminate in borderline cases (Machina 1976), denying the law of excluded middle (Field 2003), holding that vague statements can be only *relatively* true (Richard 2008), adopting a three-valued (Halldén 1949) or infinitely valued (Goguen 1969) logic, or holding that it is only true or false *to a degree* whether a vague predicate applies (Smith 2008). The *supervaluationist* tradition (Dummett 1975; Lewis 2001; Fine 1975; Keefe 2000) preserves classical validity even while rejecting classical truth-functional semantics. According to the supervaluationist, a predicate is vague when it admits of many possible *precisifications*, or ways of drawing a sharp borderline between the Fs and the non-Fs. A sentence like "A billionaire is rich" is *supertrue* because it is true on all possible precisifications, while "a person without income or assets is rich" is *superfalse*, but for the grey zone in between the sentence "someone with x net worth is rich" will be true on some precisifications and false on others. Each precisification will make (P2) false somewhere; (P2) seems intuitive because the precisification is entirely arbitrary.

This paper is not concerned with proposals which deny bivalence. I assume denying bivalence and either classical logic or semantics is an unacceptable theoretical cost (Sorensen 1994), and a more conservative approach is preferable. The most conservative way to deny (P2) is *Epistemicism.* The Epistemicist holds that there are in fact cases in which non-bald person and a bald person differ by only one hair, and that vague propositions in the grey zone are literally true or literally false, but we are simply unable to know whether propositions in this grey zone are true or false (Williamson 1994a; Sorensen 1994).

Epistemicism is widely believed to carry two burdens which make it unpalatable. First, denying the tolerance principle while maintaining bivalence is believed to conflict with our intuitions about vagueness; the tolerance intuition is thought to be analytic or part of conceptual competence (Eklund 2005; Wright 1976). Second, denying the tolerance principle while preserving classical logic and semantics is believed to entail a commitment to sharp borderlines dividing heaps from non-heaps, white from grey from black, and boys from men (Sorensen 2012; Wright 1995). The most prominent defenders of Epistemicism (Sorensen 2001; Williamson 1994b) take on both burdens.

In this paper, I argue that these are two *myths* about Epistemicism, and Epistemicism bears neither burden. I propose a Revisionary Epistemicism instead. Like other Epistemicists, I maintain that all vague propositions in the borderline region or "grey zone" are unknowable but either entirely true or entirely false. Unlike other Epistemicists, I accept a revised tolerance principle motivated by our intuitions; I simply deny that this principle is paradoxical. Unlike other Epistemicists, I entirely reject the existence of sharp borderlines; I simply deny that this requires sacrificing bivalence.

Since my Revisionary Epistemicism accepts a tolerance principle motivated by our intuitions, and denies the existence of sharp borderlines, while at the same time maintaining bivalence, it should satisfy both the conservativism about ordinary intuition which motivates bivalence-deniers and the conservativism about classical logic and semantics which motivates traditional Epistemicists. The paper will proceed as follows.

In Section 2, I reflect on the tolerance intuition. I argue that the standard *well-ordered* tolerance principle which appears in conventional Sorites arguments is not supported by intuition. What intuitions support is instead a *dense* tolerance principle, which cannot be used to generate the conventional paradox. In Section 3, I further argue that what intuition supports is a *generic* dense tolerance principle. But an Epistemicist can embrace this generic, dense tolerance principle, since it is non-paradoxical.

In Section 4, I present a “gradient” model for the metaphysics of vague predicates which is consistent with the generic dense tolerance principle. On this model, there is a fact of the matter for every vague proposition in the grey zone, yet it entails no commitment to sharp borderlines or cut off points. In Section 5, I argue that the gradient model can be justified as a semantics for vague predicates by reflection on ordinary language use.

There are other challenges which epistemicism still faces which beyond the scope of this paper. My aim is simply to revise epistemicism to appear more plausible as a response to vagueness than it seemed before, by dispelling myths which even its defenders typically accept.

2. Against Well-Ordered Tolerance

Let’s begin with some instances of (P2) from our conventional sorites:

(a) Adding 1 grain of sand to a non-heap won’t make it a heap.
(b) If you aren’t driving fast, then driving 0.01 mph faster still won’t be driving fast.
(c) If you take away 1 cent from a rich person, they’ll still be rich.
(d) If a person isn’t tall, then growing by 1cm won’t make them tall.
(e) Subtracting 1 hair won’t make a non-bald person bald.

I call these “well-ordered” tolerance principles, because the domain consists of well-ordered units of measure. Conventional sorites paradoxes rely on principles like this, which are often thought to be intuitive. While I accept that our intuitions support another kind of tolerance principle, I reject that our intuitions support tolerance principles involving well-ordered units of measure.

First, consider that well-ordered units of measure *approximate* qualitative real-world properties, which are by nature dense and perhaps continuous. For instance, the *number of years* someone has lived is a well-ordered unit of measure, which will return values like 55 or 2½, but it only approximates the real-world *age* of a person, which is dense, as one smoothly transitions between ages. Real-world *weight* is dense and continuous, but it is measured approximately in well-ordered units like “201.5 lbs.” Amounts of hair are dense, but numbers of hairs are well-ordered.

Second, well-ordered units of measure are not part of the *definitions* of vague predicates. Vague predicates are defined in terms of real-world properties, not the discrete, well-ordered units used to measure these properties. Someone is “old” in virtue of their age, not their number of years; something is “heavy” in virtue of its weight, not its number of pounds. The relationship between units of measure and the real-world property is not part of conceptual competence: one can understand “tall” without understanding inches or centimeters. The relationship between units of measure and vague predicates is thus *a posteriori*, not *a priori*. I learn that 85 decibels is loud or 50 degrees Celsius is hot through experience, not conceptual analysis. I know that light years are long and nanometers short only because I know them to be longer and shorter than miles and millimeters respectively, which I can associate with real-world quantities through experience. These are synthetic truths, not analytic truths.

So, any intuitions we have about the extensions of vague predicates must be derived from our phenomenological experiences with real-world properties, not the units of measure which approximate them. Our intuitions must be derived either from memory, perceptual experience, or–most commonly–an “armchair” act of imagination, on which one imagines a number of incremental changes and forms the judgment that no incremental change will make a difference in the application of the predicate (Wright 1976).

Fourth, both perceptual experience and armchair imagination are essentially qualitative and phenomenological, and qualitative experience does not present us with discrete, well-ordered units of measure. Phenomenal qualities present themselves in dense smudges, not well-ordered quantities. There are no phenomenal speedometers to evaluate fast and slow. What happens when I try to imagine a man who is 5 feet tall growing by one centimeter? Do I accurately imagine his particular height then increasing by one centimeter? Suppose that I imagine a measuring tape–would that make my imaginary measurements more accurate? Of course not. Rather, what I imagine is a man who isn’t too tall but isn’t too short growing by *just-a-little-bit*, some concrete, real, observable height which might or might not be best approximated by “1 centimeter.” The same is true in perception. I do not see the sun *as* falling 0.004 degrees on the horizon each second after noon.

I do not see the sun as falling at all. What I do eventually notice are some qualitative differences between perception and memory of perception, but I do not use a mental protractor to assign the difference some discrete value.

Therefore, tolerance principles about well-ordered units of measure (a)-(e) can't be supported through perceptual experience, memory, or armchair imagination. Without the aid of measuring devices, we can only consider qualitative, dense "amounts" of increase or decrease which we vaguely associate with particular well-ordered units. Imagining a non-fast car driving 0.01mph faster is like imagining a non-loud sound increasing by a microdecibel: the changes would be barely phenomenologically detectable, if they are detectable at all.

If our intuitions are formed through perception or imagination, then, our intuitions must track these claims instead:

(a*) Adding *just-a-little-bit* of sand to a non-heap won't make it a heap.

(b*) If you aren't driving fast, then driving *just-a-little-bit* faster still won't be driving fast.

(c*) If you take away a *just-a-little-bit* of money from a rich person, they'll still be rich.

(d*) If a person isn't tall, then growing by *just-a-little-bit* won't make them tall.

(e*) Subtracting *just-a-little-bit* of hair won't make a non-bald person bald.

These claims are made along the dense, qualitative domain of observation and experience. When I consider these scenarios in (a*)-(e*), I intuit that a minuscule, *just-a-little-bit* change can't make a genuine difference to the application of a vague predicate. But the scenarios I'm considering are not actually instances of the well-ordered tolerance principle (P2) in the sorites argument, but of some other, dense tolerance principle.

Therefore, our intuitions about vague predicates only support dense tolerance principles, not well-ordered tolerance principles. There is no analytic or experiential link between these intuitions and the claims about well-ordered units of measure in (a)-(e) which appear in the conventional sorites argument. Of course, through practice or inductive reasoning someone might become good at perceptually guessing discrete units of measure, like a seasoned surveyor who can visually estimate distances by the quartermeter, but these are at best heuristic and probabilistic, and not sufficient to support an exceptionless universal like (P2).

The conventional sorites is only valid if the domain is well-ordered, given restrictions on mathematical induction. So, the only tolerance principle supported by tolerance intuitions is not a principle which can be used to construct the conventional sorites argument.

3. For Generic Dense Tolerance

This only takes the Epistemicist so far, however. Recently, Weber and Colyvan (2010) have shown it is possible to take our intuitions about dense tolerance (a*)-(e*) and then generate a non-well-ordered, continuous Sorites argument within topology which dispenses with the requirement of well-ordering along a countable domain, as well as a single linear degree of variation. We can formalize a continuous tolerance principle in terms of the topology of a *connected* metric space. A metric space is connected if it cannot be partitioned into non-empty, disjoint, open sets, as is the case for the real numbers. Tolerance can be defined as the principle that, for any element in the space which has a given attribute, everything in its *vicinity* has the attribute, for some sufficiently small "vicinity." The conjunction of tolerance with connectedness leads to the conclusion that every element in the space has (or lacks) the attribute, since no sharp boundary can be applied to the application of the attribute (Dzhafarov 2019; Weber and Colyvan 2010).

Here I will present a slight modification, which I will call the *Dense Sorites*, which rests on slightly more modest assumptions (mere denseness rather than continuity), and which better conforms to the familiar pattern of reasoning in the classic well-ordered Sorites. Define a metric space M = <M,d> as *chain-connected* if and only if, for any x, y in M and $q > 0$ in $\mathbb{Q}$, there is a finite sequence of "jumps" or "steps" in the space x, z_1, z_2, ... z_n, y connecting x to y such that the distance between each step is less than or equal to q. In other words, even though the space itself may be dense and have infinitely many elements, from any point in the space we can get to any other point in the space through a finite series of jumps of any arbitrary distance q. Quantifying over elements in a metric space M = <M, d> for some distance function d(x,y), and letting $q \in \mathbb{Q}$ be our "sufficiently small" distance, we get:

(PM1) $\exists$x Fx
(PM2) $\forall y \forall z((Fy \ \& \ d(y,z) < q) \rightarrow Fz)$
(CM) $\forall$xFx

For example, letting Fx = "someone with x amount of hair is bald":

(PM1b) Someone with no hair is bald.
(PM2b) For any *amount* of hair a bald person and any *just-a-little-bit* of hair, if that *just-a-little-bit* were added or subtracted, they would still be bald.
(CMb) Anyone with any *amount* of hair is bald.

The *Strict Dense Tolerance Principle* (PM2) tells us that for some sufficiently small distance *q*, everything which is within *q* or less distance to something to which *F* applies is also something to which *F* applies. Given the chain-connectedness of our space, everything is within a finite number *n* of steps of distance *q* from everything else. So, if some element is *F*, then by a finite number of steps, we can prove for any given element of the space that it is also *F*. While the number of points in the space might be infinite, for any given point in the space there is some finite method of proving that either both are *F* or neither are, given (PM2). The property of chain-connectedness effectively replaces the role played in the conventional Sorites by well-ordering and the Archimedean axiom $(\forall x)(\forall y)(\forall z)(\exists n \in D)(x < y + (n \times z))$. This allows a Sorites argument for vague predicates which have multiple dimensions of variation of dense degrees. [1]

In Section 2, I argued that our intuitions support a dense tolerance principle. The question is now whether the principle our intuitions support is nonetheless *paradoxical*, like the *strict* dense tolerance principle (PM2). I do not know of any philosopher who defends the claim that our intuitions actually support strict dense tolerance (PM2), since most discussion focuses on well-ordered tolerance. In fact, the dense sorites has been called "a degenerate form of sorites, in which the puzzling behaviour of the classical

[1] Thanks to Roy Cook for the lesson in topology (Cook 2017).

sorites is lost precisely because the local and global level are collapsed together" (Rizza 2013). Discussions of a tolerance principle in the literature typically presuppose well-ordered tolerance. So, there is no *prima facie* reason to think strict dense tolerance is supported by intuition.

Furthermore, Susanne Bobzien has persuasively argued that people mistakenly infer strict, paradoxical tolerance principles from a weaker, modal, non-paradoxical principle (Bobzien 2025, 2597), i.e.,

> (SC2) *Weak Tolerance*, interpreting □ as "it is clear that" or "it is definite that"
> $\forall i\ \neg\Box\neg(\mathrm{F}a_i \leftrightarrow \mathrm{F}a_{i+1})$
> e.g. "for any fast speed, it is not definitely not the case that the next speed is fast"

That is, we reflect on weak tolerance whenever we form intuitions, but then invalidly *infer* strict tolerance from it. I agree with Bobzien in spirit, though I think there are a number of other non-strict tolerance principles which our intuitions and experiences could also be said to support. These principles share in common with Bobzien's (SC2) that they do not lead to a sorites paradox, and that one cannot validly infer from them the paradoxical (PM2):

> (PG2) *Generic Tolerance*, letting ⅁x be the implicit Generic quantifier in "Dogs bark"
> $\Game \mathrm{y} \Game \mathrm{z}((\mathrm{Fy}\ \&\ \mathrm{d(y,z)} < q) \rightarrow \mathrm{Fz}))$
> e.g., "Speeds a tiny bit different than a fast one are still fast."

> (PC2) *Counterfactual Tolerance*
> $\forall \mathrm{y} \forall \mathrm{z}((\mathrm{Fy}\ \&\ \mathrm{d(y,z)} < q)\ \Box\!\!\rightarrow \mathrm{Fz}))$
> e.g., "Were a speed just a tiny bit different, it would still be fast."

> (PS2) *Subjective Tolerance,* letting S(p) indicate thinking about p:
> $\forall \mathrm{y} \forall \mathrm{z}((\mathrm{Fy}\ \&\ \mathrm{d(y,z)} < q) \rightarrow (\mathrm{S(Fy}\ \&\ \mathrm{Fz)} \rightarrow \mathrm{Fz}))$
> e.g., "Any speed you'll think of as a tiny bit different than a fast one is still fast"

> (PJ2) *Doxastic Tolerance*, letting J(p) indicate justified belief in p:
> $\forall \mathrm{y} \forall \mathrm{z}(\mathrm{J(Fy}\ \&\ \mathrm{d(y,z)} < q) \rightarrow \mathrm{J(Fz)})$
> e.g., "For any speed you're justified in believing is just a tiny bit different than a fast speed, you're justified in believing it's fast too."

These four non-strict principles have in common with Bobzien's (SC2) that they could be readily justified inductively by random samplings of particular cases of vague predicates without the need to consider every possible case in an infinite, dense domain. By contrast, random sampling can't support *analytically necessary* claims about the infinite set of possible collections of wealth and possible distributions of hair, and hence does not justify an exceptionless dense tolerance principle.

These four principles also fit our intuitions about vagueness. Generic tolerance fits the phenomenology of expectation: encountering something indistinguishable from an F gives you a good reason to expect an F. Generics permit exceptions. Counterfactual tolerance fits the intuition-grabbing language of thought experiments about vagueness: *were a bald person to grow a hair, he'd still be bald.* Counterfactuals only require the truth of the *nearest* case, not every case. Neither is paradoxical.

Subjective tolerance fits the phenomenology of the *forced march*: the act of consciously considering any two qualitatively indistinguishable cases compels me to categorize neighboring cases as *F* or both as non-*F* (Horgan 1994). Since the domain is dense, this process will never force me to affirm that clear *F*s are non-*F*s, or clear non-*F*s are *F*s. Perhaps we cannot consciously consider, without falling into a kind of Blindspot, a particular case in which an F and a non-F are within close vicinity of one another (Sorensen 1988).

Doxastic tolerance best fits the phenomenology of epistemic indiscriminability. Égré (2015) proposes that, if vague predicates were not tolerant, we would expect to encounter cases in which known Fs and known non-Fs are in close vicinity to one another, but we do not. Thus, "there are no close cases in which it is known that a sentence takes a certain truth-state in one case and known that this sentence takes the complementary truth-state in the other close case" (Greenough 2003). There is no need to invoke strict dense tolerance.

Yet none of these four alternative principles are paradoxical, because each is compatible with the existence of some exceptional non-F in the vicinity of an F which disrupts the ability to form a chain of inferences throughout the space. Both individually and in conjunction with one another they are consistent with the denial of both well-ordered tolerance and of the strict dense tolerance principle, namely:

(~PM2) $\exists y\ \exists z(Fy\ \&\ d(y, z) < q\ \&\ {\sim}Fz)$

So, we can put to rest the first myth of vagueness. The myth is that our intuitions support a paradoxical tolerance principle. In fact, our intuitions do support a number of tolerance principles, but these principles cannot be used to generate a conventional sorites paradox–because their domain is not well-ordered–and they cannot be used to generate a continuous sorites or dense sorites paradox, because they admit of exceptions. Therefore, one can maintain bivalence and avoid paradox without having to deny an intuitive tolerance principle.

4. Epistemicism without Borders

The denial of well-ordered tolerance is this claim:

(~P2) $\forall c \exists x \exists y(Fx\ \&\ {\sim}F(x+y)\ \&\ |y| \leq c)$

The denial of strict dense tolerance is this claim:

(~PM2) $\forall q\ \exists y\ \exists z(Fy\ \&\ {\sim}Fz\ \&\ d(y, z) < q)$

As discussed, to avoid a Sorites, the Epistemicist must affirm both, even while they may accept other, non-strict tolerance principles. So, an Epistemicist must hold for each vague predicate that there is some point along the line where small differences impact whether or not a vague predicate applies. There is a possible bald man who differs from some possible non-bald man by only a single hair, or an even tinier amount of hair. There is a possible rich man who differs from some possible non-rich man only in owning a single penny, or some even tinier amount of wealth. This is generally interpreted to suggest that there is some sharp borderline between the bald and the non-bald, or some fixed threshold for being rich, that is:

(Borderline) $\exists x((\forall y\ (y \geq x \rightarrow Fy)\ \&\ (\forall z\ (x > z \rightarrow {\sim}Fz))$

If (Borderline) holds for richness, then there is some *poorest rich* person such that everyone richer than that person is also rich, and anyone poorer than that person isn't rich. If it holds for baldness, then there is some hairiest bald man such that everyone hairier than that bald man is non-bald and no one with less hair is bald. Of course, baldness and wealth are likely determined by some multi-dimensional dense parameter–the arrangement of the hair matters for baldness and the liquidity of the assets matters for richness–rather than a well-ordered number of hairs or dollar value in cash (Graff 2000). Still, the Epistemicist's denial of strict tolerance suggests the existence of sharp borderlines.

It does not logically *entail* the existence of borderlines, however. While it is standard to conflate the Epistemicist's claim that *there are unknowable facts in borderline cases* with the *there are unknowable borderline*s, the claim that there are sharp borderlines or thresholds is a much stronger claim. Unlike (Borderline), (~P2) and (~PM2) are consistent with the existence of a possible bald man who is hairier than some possible non-bald man, or some possible rich woman who has less wealth than some possible non-rich woman. Unlike (~P2) and (~PM2), (Borderline) requires that all possible persons who fall "between" any two possible bald persons in their degree of hairlessness must also be bald, and all possible persons who fall "between" any two possible rich persons in their degree of wealth must also be rich. Suppose that 80.492mph is the threshold for driving "fast," the least fast speed, in a given context. According to (Borderline), it follows that 80.491mph and 80.4919mph are not fast, including all of the irrational speeds between them, and that 80.49200001mph and 80.493mph are fast, as well as all the irrational speeds between them.

One could deny (Borderline), while also denying strict dense tolerance, by holding that there are some speeds less than 80.492mph which are fast, and some speeds greater than 80.492mph which are not fast; that is, 80.492 is not truly the least fast speed. Within the borderline zone, the "fast" and "non-fast" speeds might not be organized at all, but between any two speeds there may in fact be infinitely many "fast" speeds and infinitely many "non-fast" speeds.

This is a denial of a certain kind of *Monotonicity* for vague predicates. Just as we noted earlier that there are strict and non-strict forms of the tolerance principle, there are also both strict and non-strict forms of the tolerance principle, as Bobzien compares:

$\text{Monotonicity}_{\Box:}\ \forall i\ ((\Box Fa_i \rightarrow \Box Fa_{i-1}) \land (\Box\neg Fa_i \rightarrow \Box\neg Fa_{i+1}))$

$\text{Monotonicity}_{F:}\ \forall i\ ((Fa_i \rightarrow Fa_{i-1}) \land (\neg Fa_i \rightarrow \neg Fa_{i+1}))$

(Bobzien 2025, 2597, 2601).

Bobzien and I both affirm Monotonicity$_\Box$ with the □ operator interpreted as "it is clear that" or "it is definite that," or perhaps also "it is justifiable to believe that" or "I actively consider that" or "it is accepted in conversation that." But Bobzien raises doubts about Monotonicity$_F$:

> There is some empirical evidence that even qualified speakers seem inconstant and capricious in their assessment of such objects with regard to F. They may judge ak+n to be F but ak+(n−1) not to be F, and even may judge the same object first F and shortly after ¬F, if they are unaware that it is the same—or even if they are cognizant of this fact (e.g. (Raffman 1994) (Bobzien 2025, 2601).

Similar concerns about Monotonicity$_F$ given the multi-dimensionality of vague predicates are raised by Sagid Salles:

> Not just the number of hairs, but also how they are distributed on someone's head is relevant to the application of this predicate. Because of this, it is arguable that a person who has n hairs on her head is bald, while a person who has fewer than n hairs is not bald. We could handle this problem by saying that, all other things being equal, $\forall n\ (Ba_n \rightarrow Ba_{n-1})$ will hold. (Salles 2021, 133)

An Epistemicist can maintain bivalence while rejecting (Borderline) by accepting the weaker Monotonicity$_\Box$ while rejecting the stricter Monotonicity$_F$. I hope an illustration will help show why this is actually a very good picture of what the extension of vague predicates looks like.

Imagine that you are in a room facing a large wall. The wallpaper on the wall creates a black-to-white gradient. On the left side, the wall is completely black, and on the right side, the wall is completely white. The wall smoothly transitions from black to white with every shade of grey in between.

Figure 1

Clearly, there is no “borderline” or “threshold” point along the wall *t* such that everything to the left of *t* is black and everything to the right of *t* is not black.

Suppose, though, that the wall-paper was printed using a dot-matrix printer from the 1980s. For those too young to recall, a dot-matrix printer would print only tiny black dots and white dots, although when viewed at the level of an ordinary human eye the aggregate of individual black and white dots appeared gray. Then it *will* be true that there exist points *p* which are black, but that points immediately to their right (and their left!) are white. Such points will be scattered all over the “grey” region at the center of the wall. This doesn’t make *p* a borderline or threshold, but does make well-ordered tolerance false.

Suppose now that this is a futuristic dot-matrix printer, which prints using infinitesimally small points. The result is a *perfect* gradient along a continuum. No matter how powerful a magnifying glass one takes to the wallpaper, one will never see the individual dots. There will be nothing remotely resembling a borderline. It will be true, however, that for any well-ordered distance to the right *n*, there is some point on the wall *p* such that *p* is black, but *p+n* is not black – and it will be false that there is a sufficiently small well-ordered distance to the right *n* such that for all *p*, if *p* is black, then *p+n* is black. Both strict dense tolerance and (Borderline) will be false.

In some regions of the wall, the relative distribution of white points to black points will be 99-100% or 0-1%. Suppose I were to throw an infinitesimally small-tipped dart at one of those regions of the wall. Then “I will hit a black spot” or “I will hit a white spot” will both be highly probable and assertable. In the middle of the wall, the distribution will be 50/50, and “I will hit a black spot” will be unknowable and unassertable. In other regions the distribution may be 95% or 90% or 85%–at some point “I will hit a black spot” will pass from assertable to unassertable, from knowable to unknowable. But this does not mean that it becomes indeterminate whether I will hit a black spot or a white spot, nor does it mean that it is 95%, 90%, or 85% *true* that I will hit a black spot. It is not 90% true that the region is black, but rather it is 100% true that 90% of the region is black.

On this picture, the facts in borderline cases are stochastic, random, chaotic, erratic, and unpredictable, just like throwing a dart at the grey zone of the gradient on the wall. They are stochastic because meaning is grounded in actual use, and the actual use of vague predicates is stochastic in the borderline range of cases. There will be an objective probability that the dart will hit a white spot instead of a black spot–a probability which will increase to the right and decrease to the left–but *only* a probability. We can never know. Similarly, in borderline cases of *heap* or *bald*, a speaker will have some probability of speaking the truth, but cannot *know* that what she is about to say is true, since every borderline case of baldness is surrounded on all sides by possible borderline cases of non-baldness. The appropriate thing to do, given the knowledge norm of assertion and Gricean principles, is to say something hedged in the way ordinary speakers do.

Notice that this model of vagueness allows our four non-strict tolerance principles to still come out true. *Generic Tolerance* will still be true, because in general it is true that Fs will be generally surrounded by other Fs. Even the stray Fs which are in the region of the borderline zone closer to the non-Fs will have other Fs nearby them. *Doxastic Tolerance* will be true for the same reason, given that the probability of running into an F in the vicinity of another F is going to remain high. These principles require only something like Monotonicity$_{\Box}$, not Monotonicity$_{\mathrm{F}}$.

The principles of *Subjective Tolerance* and *Counterfactual Tolerance* will come out true if we specify some further rules about context-shifting and reference. Let's specify that an accepted utterance of "n is an F" impacts every subsequent utterance of "$n+m$ is an F" by "magnetically" pulling uses of "$n+m$" to refer to a particular point the region of our *M*-space which is an F as opposed to one which is not an F, *ceteris paribus.* For instance, suppose that it is accepted in a conversation that "56mph is fast," and then I utter "55mph is fast." There are infinitely many particular speeds within the range of 54.5 - 55.5mph to which "55mph" could refer. Suppose that 55mph is towards the "non-fast" end of the gradient in the borderline zone for "fast," such that 20% of the points are "fast" and 80% of the points are "non-fast." While it would be more probable that we'd land on a "non-fast" point were we to jump to a random point within the 55mph range, the pull from the acceptance of "56mph is fast" will lead "55mph" to refer to one of the 20% of the particular speeds which *are* fast, rather than one of the 80% of particular speeds which are not fast. So, “55mph is fast” will come out as true. If we grant that this context-sensitive reference magnetism impacts which cases I am able to think about–given that my mind cannot phenomenologically distinguish small differences between speed, which particular speed my thought “55mph” refers to is partly fixed by “56mph is fast” – then *Subjective Tolerance* will be true. If our analysis of counterfactual nearness includes such context sensitivity, then it will also come out true that “Were you to drive 1mph slower, you’d still be driving fast.”

My revisionary Epistemicism can affirm a number of other related “anti-borderline” claims which a traditional Epistemicist would have to deny:

> (No Carvings) It isn’t possible to carve up the extension of F into two disjoint non-empty open subsets such that everything in one is F and nothing in the other is F, or such that everything connected to one side is F and everything connected to the other side is not F.
>
> (No Crossings) Take any point in the extension, and head in any direction from it along a path. There’s no point such that prior to that point there are Fs and only Fs, and after that point there are non-Fs and only non-Fs.
>
> (No Zooming) There will be regions R such that every sub-region of R will have the same distribution of Fs and non-Fs as R. In other words, one can zoom in far enough that zooming in on the grey further will no longer change the proportion of black and white, but it will be a consistent shade of grey.
>
> (Infinite Buffers) For anything which is F, there’s something else which is F and which is *closer than* the nearest thing which is not F, and vice-versa.
>
> (No Least Greatests) For anything which is F, there may be something else which is F and which is *closer to* the nearest thing which is not F. Hence, there is no shortest possible tall person, because for any tall person, however short, there could be a shorter tall person.

> (Borderline-Borderline Zones) Between any two non-empty regions R_m and R_n where the distributions of Fs are $n\%$ and $m\%$ respectively, there will be regions in between them where the distribution of Fs is between n and m. Further, it is not possible to divide any region R into two disjoint non-empty open subsets such that for one the distribution is 0% and for the other the distribution is >0%. In other words, there will always be borderline zones as to whether a region is a borderline zone.
>
> (Lotteries) In the grey zone, while I rationally must believe that Fx and ~Fy for some neighboring x and y, yet for each particular neighboring x and y I rationally must believe that both are F or both are not F. The sorites is thus analogous to the lottery paradox (Lissia 2022)

I take these to be characteristics of genuine, *robust* vagueness. My theory is thus not a "wimpy" theory, to use Terry Horgan's term, as it invokes no arbitrary borderlines or precisifications. I escape the "problem of precisification" (Horgan 1994), the problem of the arbitrariness imposing constraints that sharpen the extension, because I entirely deny anything associated with sharpenings. Yet, unlike Horgan, I do not have to reject bivalence or the determinate truth of vague propositions. I simply have to make the domain dense instead of well-ordered, and to scatter the extension and antiextension across a gradient.

I might appear to face, however, something analogous to what Horgan calls the "foundational problem of precisification" faced by supervaluationists. That is, *what grounds* the fact that x is F for some borderline case of F-ness? On my account, why is one borderline case of baldness truly bald, even though two other possible men with indistinguishable amounts of hair aren't bald? What kind of strange truthmaker could make "n kmph is fast" true, and "$n + \pi^{-3}$ kmph is fast" false?

5. Grounding the Scattered Model

Meaning is grounded in use. In practical, everyday life, the use of vague predicates is stochastic and chaotic. Given this, there are only three possible attitudes to take towards vague predicates:

(a) Use *underdetermines* meaning
(b) Use *determines* meaning *in an orderly, precise way*
(c) Use *determines* meaning *in a stochastic, chaotic way*

The bivalence-deniers take attitude (a). For instance, they might say that vague predicates suffer from an incompleteness of meaning (Fine 1975). Perhaps they are akin to indexicals in having character but not content until assigned a boundary in a local context and situation: the meaning of “heavy” is only fixed when we have to divide up the actual wrestlers evenly into “heavyweights,” “middleweights,” and “lightweights” (Salles 2021, 151). Perhaps this is akin to grading papers, where the meaning of a “B+” shifts based on who is in the class and where one feels like drawing the cut off (Maudlin 2008). Perhaps use determines meaning in clear cases, but for borderline cases use fails to determine a truth value (Tye 1994). Alternatively, perhaps use determines a meaning which is genuinely inconsistent and paradoxical (Eklund 2005).

Traditional Epistemicists take attitude (b). They hold that precise meanings can emerge from patterns of use, even though these boundaries of these precise meanings are unknowable. Tolerance is an illusion produced by the persistence heuristic (Williamson 2024, 1.3). Sharp thresholds arise from collective behavior, though their location is as unpredictable and ever-shifting as a weather forecast (Williamson 1994b).

My revisionary epistemicism takes attitude (c). Given that meaning is grounded in use, and the use of vague predicates is chaotic and stochastic, then – even controlling for local context–we should expect the extension of a vague predicate to be fully determined in an equally stochastic and chaotic way. This position is more conservative than (a), while avoiding the implausible commitments of (b).

If it helps to get a grip on the truthmakers for vague propositions on my view, imagine that in some Platonic heaven a zealous prosecutor refers every possible thought with a vague concept and every possible utterance with a vague term before a jury of ideally competent speakers, along with a detailed review of the context and a description of the various basal properties, both those known to the conceiver or speaker and those unknown, and a review of the precedents of accepted and rejected past uses of the predicate. The jury will then deliberate. If it comes to a consensus that the proposition is false, then the proposition is false. If it fails to achieve a consensus that it is false, then the proposition is true. The truthmakers will thus include the multi-dimensional basal properties, the context, and the social facts about use, but also something like *how the idealized judgments happen to fall.*

We should expect judgments in these cases to fall predictably and reliably in clear cases, and chaotically and randomly in the grey zone. Some possible borderline old men will be just slightly older than possible borderline non-old men, and some just slightly older, and some possible men of the same age will be old while others are not. Monotonicity$_F$ will fail. Nonetheless, the consensus or non-consensus of the jurors will settle the matter.

Although I have spoken of a jury, the relevant point is not *subjectivity* but arbitrariness. Consider Crispin Wright's "tachometer paradox." A tachometer measures the rotational speed of the motor, and produces a well-ordered numerical output in RPMs given the constantly varying and dense real-world state of the motor. There are inevitably borderline cases in which a miniscule change in the motor's rotation leads to a shift up or down in the tachometer in one case, and not in another case. The shifts will be arbitrary, but no human subjective judgment will be involved. Stochasticity is no reason to doubt there is a fact of the matter (Wright 1987).

When we consider the multi-dimensional, context-sensitive properties which ground baldness, along with the stochasticity of ordinary use of "bald," it becomes very hard to believe the extension of "baldness" will be better behaved than the reading of the tachometer. Even *purely hypothetical* sharp borderlines, or the idealized possible "precisifications" or "sharpenings" proposed by the supervaluationists, seem at odds with the meanings of vague terms. Instead, we should expect true uses of any vague term–say, "is brave"–to cluster stochastically and non-monotonically around a nucleus, not unlike an electron cloud:

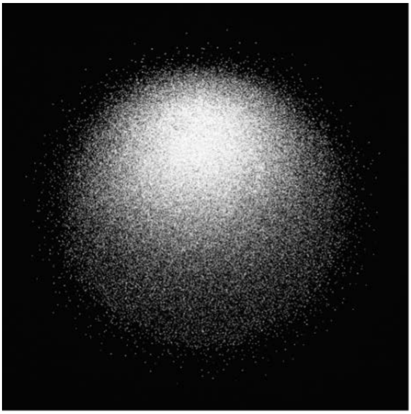

Figure 2

This model of the extension of a vague predicate affirms that “vagueness, if taken at face value, cannot be reconciled with *any* precise dividing lines.” (Tye 1994, 193) Surprisingly, it shows that it is consistent to hold that there are no sharp borderlines between the extension and antiextension for *P*, and yet every object is either in the extension or antiextension of *P*. This in itself is sufficient to dispel the second myth of vagueness, and to lift the second burden from epistemicism. The revisionary Epistemicist can accept that it is analytic that there are no shortest possible tall people or first noonish moments without any denial of bivalence.

6. Conclusion

I have presented a revisionary form of epistemicism which affirms its essential doctrine: every vague proposition is either true or false, even if it is unknowable in the borderline region. In comparison to supervaluationism and degree of truth views, my model remains conservative by preserving the bivalence of classical logic and semantics. On the other hand, unlike traditional epistemicism, my model allows us to affirm that there are no strict borderlines, thresholds, or cut off points between the bald and the non-bald, because the extensions of vague predicates are scattered stochastically and non-monotonically in the borderline region, akin to the readings of Wright’s tachometer–just like we should expect, given the dependence of meaning on use.

My model also allows us to affirm a number of serious tolerance principles, supported by ordinary intuitions, by distinguishing these from two paradox-generating tolerance principles which I have argued are not backed by intuition. Our intuitions must be formed through experience, but our experiences are too qualitative and *dense* to support the well-ordered tolerance principle historically used to generate sorites paradoxes. While a dense sorites may be generated by a strict dense tolerance principle, intuitions do not extend so infinitesimally as to support such a principle. It is more plausible to hold, like Bobzien (2025), that we form intuitions that support a weak or non-strict principle, and then make mistaken inferences to stricter principles.

I have only discussed the semantics of vagueness. I admit my revisionary epistemicism needs still to offer an account in pragmatics for why speakers accept "he didn't plant the petunias, but he didn't *not* plant them" (Machina 1976), or why social pressure in a forced march can stretch what speakers are willing to accept as the extension of a vague predicate. I have not argued explicitly against supervaluationist or degree of truth proposals; sacrificing bivalence may be rational if one thinks epistemicism is committed to affirming sharp borderlines and rejecting the tolerance intuition, much as Williamson and Sorensen present it. Insofar as I have offered a way in which bivalence can be preserved along with the denial of sharp borderlines and a charitable interpretation of the tolerance intuition, however, I take my project of demythologizing to have reduced the appeal of these views.

Acknowledgements

I am indebted to comments on early drafts from Roy Cook, Geoff Georgi, and audiences from presentations at the American Philosophical Association Central Division meeting (2017), and the Pittsburgh Area Philosophy Colloquium (2016), along with several helpful anonymous referees. I also have appreciated discussions on the topics in this paper with Bernard W. Kobes, David McElhoes, Steven Reynolds, Ángel Pinillos, Jacob Gittins, Jason Hanschmann, and Klaus Ladstaetter.

References

Bobzien, Susanne. 2025. "A Generic Solution to the Sorites Paradox." *Erkenntnis* 90(6): 2593–2632. https://doi.org/10.1007/s10670-023-00731-1

Cook, Roy. 2017. "Comments, Colloquium: Philosophy of Language." Handout, American Philosophical Association Central Division Meeting, Kansas City, Missouri.

Dummett, Michael. 1975. "Wang's Paradox." *Synthese*, 30(3–4): 201–32. https://doi.org/10.1007/bf00485048

Dzhafarov, Damir D. 2019. "A Note on the Reverse Mathematics of the Sorites." *Review of Symbolic Logic*, 12(1): 30–36. https://doi.org/10.1017/s1755020318000461

Égré, Paul. 2015. "Vagueness: Why Do We Believe in Tolerance?" *Journal of Philosophical Logic*, 44(6): 663–79. https://doi.org/10.1007/s10992-015-9352-z

Eklund, Matti. 2005. “What Vagueness Consists In.” *Philosophical Studies*, 125(1): 27–60. https://doi.org/10.1007/s11098-005-7773-1

Field, Hartry. 2003. “No Fact of the Matter.” *Australasian Journal of Philosophy*, 81(4): 457–80. https://doi.org/10.1080/713659756

Fine, Kit. 1975. “Vagueness, Truth and Logic.” *Synthese*, 30(3–4): 265–300. https://doi.org/10.1007/bf00485047

Goguen, J. A. 1969. “The Logic of Inexact Concepts.” *Synthese*, 19(3–4): 325–73. https://doi.org/10.1007/bf00485654

Graff, Delia. 2000. “Shifting Sands.” *Philosophical Topics*, 28(1): 45–81. https://doi.org/10.5840/philtopics20002816

Greenough, Patrick. 2003. “Vagueness: A Minimal Theory.” *Mind*, 112(446): 235–81. https://doi.org/10.1093/mind/112.446.235

Halldén, Sören. 1949. *The Logic of Nonsense*. Uppsala: Uppsala Universitets Arsskrift.

Horgan, Terence. 1994. “Robust Vagueness and the Forced-March Sorites Paradox.” *Philosophical Perspectives*, 8: 159–88. https://doi.org/10.2307/2214169

Keefe, Rosanna. 2000. *Theories of Vagueness*. Cambridge: Cambridge University Press.

Keefe, Rosanna. 2007. “Vagueness Without Context Change.” *Mind*, 116(462): 275–92. https://doi.org/10.1093/mind/fzm275

Lewis, David K. 2001. *On the Plurality of Worlds*. Oxford: Blackwell Publishers.

Lissia, Lina Maria. 2022. “Cut-off Points for the Rational Believer.” *Synthese*, 200(2): 98. https://doi.org/10.1007/s11229-022-03510-7

Machina, Kenton F. 1976. “Truth, Belief, and Vagueness.” *Journal of Philosophical Logic*, 5(1): 47–78. https://doi.org/10.1007/BF00263657

Maudlin, Tim. 2008. “Grading, Sorting, and the Sorites.” In *Midwest Studies in Philosophy*, edited by Felicia Ackerman. Minneapolis: University of Minnesota Press. https://doi.org/10.1111/j.1475-4975.2008.00170.x

Raffman, Diana. 1994. “Vagueness Without Paradox.” *Philosophical Review*, 103(1): 41–74. https://doi.org/10.2307/2185872

Raffman, Diana. 2005a. “Borderline Cases and Bivalence.” *Philosophical Review* 114(1): 1–31. https://doi.org/10.1215/00318108-114-1-1

Raffman, Diana. 2005b. “How to Understand Contextualism About Vagueness: Reply to Stanley.” *Analysis*, 65(3): 244–48. https://doi.org/10.1111/j.1467-8284.2005.00558.x

Richard, Mark. 2008. *When Truth Gives Out*. Oxford: Oxford University Press.

Rizza, D. 2013. “Deconstructing a Topological Sorites.” *Philosophia Mathematica*, 21(3): 361–64. https://doi.org/10.1093/philmat/nkt025

Salles, Sagid. 2021. “Vagueness as Arbitrariness.” In *Vagueness as Arbitrariness: Outline of a Theory of Vagueness*, edited by Sagid Salles. Cham: Springer.

Smith, Nicholas J. J. 2008. *Vagueness and Degrees of Truth*. Oxford: Oxford University Press.

Soames, Scott. 1998. “Vagueness, Partiality, and the Sorites Paradox.” In *Understanding Truth*, edited by Scott Soames. Oxford: Oxford University Press. http://doi.org/10.1093/0195123352.003.0008

Sorensen, Roy. 2012. “Vagueness.” The Stanford Encyclopedia of Philosophy (Summer 2012 Edition), edited by Edward N. Zalta. Available online at: https://plato.stanford.edu/archives/sum2012/entries/vagueness/

Sorensen, Roy A. 1988. *Blindspots*. Oxford: Oxford University Press.

Sorensen, Roy A. 1994. “Symposium: Vagueness and Sharp Boundaries.” *Mind*, 103(409): 47–54. https://doi.org/10.1093/mind/103.409.47

Sorensen, Roy A. 2001. *Vagueness and Contradiction*. Oxford: Oxford University Press.

Stanley, Jason. 2003. “Context, Interest Relativity and the Sorites.” *Analysis*, 63(4): 269–81. https://doi.org/10.1111/1467-8284.00436

Tye, Michael. 1994. “Sorites Paradoxes and the Semantics of Vagueness.” *Philosophical Perspectives*, 8: 189–206. https://doi.org/10.2307/2214170

Weber, Zach, and Mark Colyvan. 2010. “A Topological Sorites.” *Journal of Philosophy*, 107(6): 311–25. https://doi.org/10.5840/jphil2010107624

Williamson, Timothy. 1994a. “Vagueness.” *British Journal for the Philosophy of Science*, 46(4): 589–601. https://doi.org/10.4324/9780203014264

Williamson, Timothy. 1994b. *Vagueness*. London: Routledge.

Williamson, Timothy. 2024. “Heuristics.” In *Overfitting and Heuristics in Philosophy*, edited by Timothy Williamson. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780197779217.003.0001

Wright, Crispin. 1976. “Language Mastery and the Sorites Paradox.” In *Truth and Meaning: Essays in Semantics*, edited by John McDowell and Gareth Evans. Oxford: Clarendon Press.

Wright, Crispin. 1987. “Further Reflections on the Sorites Paradox.” *Philosophical Topics*, 15(1): 227–90. https://doi.org/10.5840/philtopics198715118

Wright, Crispin. 1995. “The Epistemic Conception of Vagueness.” *Southern Journal of Philosophy*, 33(S1): 133–60. https://doi.org/10.1111/j.2041-6962.1995.tb00767.x

Organon F 33 (1) 2026: 143–150 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2026.33106

BOOK REVIEW

Finnur Dellsén: *Abductive Reasoning in Science* Cambridge: Cambridge University Press, 2024, 84 pages

Liliana Bokroš*

Finnur Dellsén's book, even though narrow, but very rich in content, was created as a part of the Cambridge's series: Cambridge Elements in Philosophy of Science. These Elements are constituted by various topics from the Philosophy and Methodology of Science.

Dellsén's Element consists of four chapters, across which he presents the readers with the main approaches to abductive reasoning in science. He does not leave some of the most influential and discussed objections and challenges aside. Furthermore, he makes the matter far more accessible by providing the reader with examples and helpful classifications.

The first chapter, *A Brief History of Abductive Reasoning*, is dedicated to track the origin of abductive reasoning not just to Charles S. Peirce, but even further, to the work of Francis Bacon and René Descartes. He shows that the abductive reasoning firstly emerged as a scientific practice rather than an explicit theory. Presenting the historical cases from such figures of modern science as Antoine Lavoisier and Charles Darwin, he supports this claim, since the structure of abductive reasoning can be observed as implicit within their most influential research. He continues this history cruise by presenting Peirce's schema of abduction, with the important note, that Peirce's use of the term "abduction" is distinct from how it is used and understood these days. For someone, who takes this Element to be an introduction to abductive inferences,

* Comenius University in Bratislava
https://orcid.org/0009-0002-8013-4000
Department of Logic and the Methodology of Sciences, Faculty of Arts, Gondova 2, 814 99 Bratislava, Slovak Republic
liliana.bokros@uniba.sk

this note is important, since the terms “abduction” and “inference to the best explanation” are often used synonymously and interchangeably in the relevant literature.

From Peirce, Dellsén moves to hypothetico-deductive (HD) model, which can be seen as another forerunner for the contemporary abductive reasoning. The difference between HD model and Peircean abduction lies in the main idea behind each account, that is, for abduction (Peircean) the goal is to formulate new theories, but for HD model (as presented by Hempel) it is hypothesis evaluation. Then the similarity, which is based on their structures, since “both require a kind of derivation of a manifest fact from a hypothetical guess” (p. 9), is even more curious because of above mentioned difference. The problem which arises with both accounts is that for a proposition, “surprising fact C” (in Peircean abduction) or “empirical consequences E” (in Hempel’s HD model), there might be various incompatible propositions that make it a matter of fact (or theories from which E may be derived in HD model).

Last section of his historical tour is devoted to Gilbert Harman’s IBE – Inference to the Best Explanation. Harman’s main idea of such inference is that the truth of the hypothesis is inferred upon its capability to provide a better explanation of evidence, than any other would. This kind of reasoning is purely comparative. Dellsén then shows how Harman’s IBE is an improvement when compared to HD model. Many abductive inferential accounts are these days labeled as IBE whether they are improvements of Harman’s or not.

With this, Dellsén moves to the second chapter of the Element – *Contemporary Accounts of Abductive Reasoning.* He distinguishes between *inferential, probabilistic* and *hybrid* accounts.

Inferential accounts consist of inferring hypotheses based on their explanatory considerations. Dellsén states that in this sense, most accounts of abductive reasoning are inferential. It can be found for example in Harman’s (1965), but also in Lipton’s (1991). Lipton (1991) is explicit on what constitutes the *best* or *better* explanation – it is the one that would be the *loveliest*, i.e. it would provide the most understanding, if it were true. Some issues may arise with inferential accounts because their evaluative structure often differs. Since it is the explanatory virtues (such as simplicity, scope, unification etc.) that determine the explanatory goodness, one compares the degree of virtuousness of one hypothesis with the rest of those already formulated, but not with other potential ones.

Probabilistic accounts on the other hand situate the abductive reasoning into the Bayesian framework. The crucial part of Bayesianism for abductive reasoning is the Bayes' theorem which shows how is the posterior probability dependent on the prior probability, likelihood and expectedness of evidence, and the Rule of Conditionalization that tells a rational agent how to update their credences after obtaining some new evidence E.

One of the probabilistic accounts that Dellsén presents is from van Fraassen's critique of inferential accounts. The idea is that if abductive reasoning is to be implemented in Bayesian framework, one should award the best explaining hypothesis greater posterior probability (probability of hypothesis conditional on the evidence) than Bayesianism would. Dellsén calls this conditionalization with explanatory bonus an *Abductive Conditionalization*, since the explanation plays its role. However, van Fraassen himself claims that this is not plausible, because it is in conflict with the original Rule of Conditionalization.

Another possibility is to let explanatory consideration constrain Bayesianism, specifically by influencing the probability values that are at its core. This would then save Bayesian Conditionalization, but it would not justify the abductive reasoning, since there is no rational ground for such constraints to be required.

Last group of abductive accounts Dellsén calls hybrid, since they are a combination of inferential and probabilistic features of abductive reasoning. The *dualistic* hybrid account can be understood as a simultaneous and independent occurrence of two distinct forms of reasoning that can both be influenced by explanatory considerations in a single agent. The *heuristic* account on the other hand takes one type of reasoning to be more normatively fundamental, and mostly it is the probabilistic reasoning. For such heuristic to be plausible, the abductive inference (specifically a standard form of IBE) ought to approximate normatively correct probabilistic reasoning, which could be expected, since explanatory considerations have an impact on both accounts. Hence, heuristic accounts "must arguably assume that rational probability assignments favor hypotheses that provide better explanations, as per probabilistic accounts" (p. 29).

Such overview of various abductive accounts with their advantages and disadvantages makes it much easier for one to compare and evaluate these accounts. And if this Element would be used as an introduction to the issue of abductive reasoning, probabilistic and hybrid accounts show right away that there are some other approaches to scientific reasoning, with which combining the abductive approach can have a great impact.

In the third chapter, Dellsén aims to eliminate any doubts about whether or why explanatory hypotheses should be preferred. Since abductive reasoning is based on inferring a higher probability to a hypothesis that best explains some evidence, the most important part is to make clear what does it mean to be a best explaining hypothesis – we have a question about what constitutes explanatory goodness (or power, as it is often referred to). Dellsén presents two conceptions of explanatory goodness: *virtue-theoretic* and *subjunctive* conception.

Virtue-theoretic conception is based on a list of theoretical features called explanatory virtues that a hypothesis instantiates. But what counts as such virtue? In the literature these days we can find lots of lists that (from author to author) vary in quantity, classification and explication of these virtues, but it is possible to identify some of them that are most cited (i.e. Dellsén refers to scope, parsimony, unification, plausibility and analogy). One of the most discussed problems with virtue-theoretic conception is the non-measurability of the virtues. Though, some philosophers came with probabilistic measures of explanatory power or goodness, which represents the degree of a virtue in quantitative terms, there still cannot be found any consensus on how to understand or explicate the explanatory goodness. Subjunctive conception, with Elliot's (2021) terminology, covers Lipton's (2004) idea of "loveliness" which is determined by how much understanding is provided by a hypothesis, supposing it is true. Dellsén then points out one detail – that Lipton is noncommittal about what counts as an explanatory virtue and quite thrifty on the question of what constitutes the *loveliness* of hypothesis. This shows that the subjunctive conception needs an elaboration.

The next chapter focuses on how the explanatory goodness relates to truth. If a better explanation of a hypothesis means that it is more virtuous, and if those virtues are epistemic, then a better explanation should also mean that the hypothesis is more likely to be true. Dellsén calls this view *realism about explanatory goodness.* He also discusses the opposite view – *antirealism about explanatory goodness* – the idea that the better explaining hypotheses are not more likely to be true than those which are not as good in explaining. Dellsén then ends this classification with a note that it is in fact reasonable to take some of the explanatory virtues as truth-conducive and others as not. Furthermore, he demonstrates the difference in regard of explanatory goodness between realists and antirealists on two particular virtues (*scope* and *parsimony*). To conclude this section, Dellsén points out another, more plausible view of the explanatory goodness, which he calls *contextualism* – theory's truth-conduciveness is context-dependent, i.e. it depends on other factors that are beyond the theory or hypothesis itself.

The last chapter titled *Is Abductive Reasoning Irrational?* is focused on the main objections and challenges that abductive reasoning faces, most of which are based on questioning its rationality.

Perhaps the most influential and discussed objections to a form of abductive reasoning, come from Bas van Fraassen (1989, 142–143). The *bad lot objection* to IBE basically points out, that we cannot ever be sure that the best hypothesis is among those considered, hence we might find ourselves choosing the best out of a bad lot. Even though, many authors have tried to show that van Fraassen is mistaken, or that his objection has no merit, Dellsén showed some of the historical examples where scientists in fact found themselves choosing a hypothesis from a bad lot. He then shows possible responses to this objection, which he divides to *reactionary* and *revisionary* responses. Every one of them is then accompanied by a rejoinder which could disrupt the original response.

Such influential response might be found in Lipton (1993), who identifies two stages of IBE – the formulation of rival explanatory hypotheses and comparative evaluation of those hypotheses. According to Lipton, if someone is reliably capable of the latter, they will also be reliably capable of the former – that is because a reliable comparative evaluation needs to be based on a large set of true background theories, that themselves would have been generated by IBE in an earlier time.

Revisionary responses to the bad lot objection hold that such objection should be a reason to reformulate or replace IBE with some other account of abductive reasoning, as it may help to avoid it. Such responses can be found in Kuipers (2000) or Dellsén (2018) – they are based on the development of an abductive account which does not guarantee an absolute conclusion about the hypothesis. Rather they use a comparative conclusion saying that it is closer to the truth (Kuipers, ibid) or more likely to be true (Dellsén, ibid), than its competitors. Then, for the cases in which IBE can effectively establish the absolute conclusion, e.g. Musgrave (1988) and Lipton (2004) suggested a modification of IBE, as to include a requirement for the inferred hypothesis not only to be better than others available, but also to be *good enough*, sufficiently explanatory to be inferred at all.

In this light Dellsén (2021) developed a different account for the justification of inferring a good enough hypothesis, which is based on a process of *explanatory consolidation.* It includes accommodation of two different types of information which would make the considered hypothesis more plausible as better explanation of the evidence at hand than any other possible hypothesis.

The second van Fraassen's objection is known as *The Dynamic Dutch Book Argument.* Van Fraassen uses this argument as a motivation for moving from inferential accounts (such the Harman's and Lipton's) to a sort of probabilistic one, which he then ends up also rejecting. His objection is pointed on the earlier mentioned *Abductive Conditionalization,* specifically saying that it conflicts with Bayesian Conditionalization, since it requires assigning bonus points to a hypothesis, which according to Bayesian Conditionalization it does not deserve. And since any updating of probabilities that conflicts with Bayesian Conditionalization guarantees one to lose in a bet from a Dutch bookie, and an agent knows that this is the risk of following any other rule (Abductive Conditionalization in this case), van Fraassen claims, that such agent is *diachronically incoherent*, therefore irrational. Dellsén then presents Igor Douven's version of Abductive Conditionalization in which not only bonus points are added to the best explanatory hypothesis, but also penalty (zero points) to all the competitors. Even if this move does not completely avoid the vulnerability of a Dutch bookie, it is not irrational to use this form of conditionalization, since in some cases it might be even more beneficial than the Bayesian account.

Lastly in this section Dellsén presents some of the most recent challenges to abductive reasoning – *The Screening-Off Challenge, The Problem of Multiple Rivals* and *Incoherence Across Explanatory Levels.*

The screening-off challenge is focused on the abductive reasoning within the Bayesian framework, especially on the role of the explanatory considerations. Sober and Roche (2013) argue that the fact, that if a hypothesis were to explain some evidence, it should raise the probability of such hypothesis more than would evidence itself, can be seen as *evidentially irrelevant,* according to their *screening-off criterion.*[1]

The problem of multiple rivals depicts that the result of abductive reasoning is the one hypothesis that does best in explaining concerned evidence regardless of the fact, that among considered hypotheses there may be several plausible explanations, that are nearly as good as the best one. Dellsén's (2017) suggestion is that within the inferential and hybrid accounts this problem can be faced after a generalization of IBE to *abductively robust inference (ARI).* What ARI gives us is the possibility of inferring a claim that is entailed by several hypotheses (set of best explanations for that claim), from which at least one would be true.

Last of the recent challenges is the Incoherence across explanatory levels, which aims at the fact that a phenomenon can be explained at multiple levels. So, if we were to infer various hypotheses from the same evidence, those hypotheses do not necessarily all have to be mutually competing. Rivals would be only those which are at the same level. Climenhaga (2017) suggests that some of abductive accounts would then be incoherent, since this would allow to make inferences that are incompatible with each other. Dellsén then suggests that there is a privileged level of explanation for IBE to operate on, this level would be that "at which the set of hypotheses provide more informative explanations" (p. 62).

To conclude this abductive journey, there is not much I can add to. Dellsén's work on this element is perfectly balanced – he goes into details of some particular cases or accounts where it is the case that a fine detail makes a difference. On the other hand, he does not do so in passages where so many details would do nothing more than confuse the reader. That is the main reason why this Element is a great starter point for any beginner in the field of abductive reasoning. Furthermore, his classifications serve not only to better understand individual accounts of abductive reasoning or approaches to explanatory goodness, but also to easily navigate through the text, especially in cases where he refers to previous passages. We could say that Dellsén, in just a few pages, managed to introduce the basics of abductive reasoning in science, supporting many of them with examples, but also pointed out important objections that many authors have to the use of IBE in science, as well as potential answers

[1] Explanatory considerations are evidentially irrelevant if Pr(H|E&X) = Pr(H|E), where X is the fact that H explains E. In such case one could say that E screens off H from X.

and open questions related to them. This Element therefore stands as a successful and concise guide to abductive thinking, accessible to everyone.

References

Climenhaga, Nevin. 2017. “Inference to the Best Explanation Made Incoherent.” *Journal of Philosophy*, 114(5): 251–73. https://doi.org/10.5840/jphil2017114519

Dellsén, Finnur. 2017. “Abductively Robust Inference.” *Analysis*, 77, 20–29. https://doi.org/10.1093/analys/anx049

Dellsén, Finnur. 2018. “The Heuristic Conception of Inference to the Best Explanation.” *Philosophical Studies*, 175: 1745–66. https://doi.org/10.1007/s11098-017-0933-2

Dellsén, Finnur. 2021. “Explanatory Consolidation: From ‘Best’ to ‘Good Enough.’” *Philosophy and Phenomenological Research*, 103: 157–77. https://doi.org/10.1111/phpr.12706

Dellsén, Finnur. 2024. *Abductive Reasoning in Science*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009353199

Elliott, Katrina. 2021. “Inference to the Best Explanation and the New Size Elitism.” Philo*sophical Perspectives*, 35: 170–88. https://doi.org/10.1111/phpe.12148

Harman, Gilbert. 1965. “The Inference to the Best Explanation.” *The Philosophical Review*, 74: 88–95. https://doi.org/10.2307/2183532

Kuipers, Theodorus A. F. 2000. *From Instrumentalism to Constructive Empiricism*. Dordrecht: Springer. https://doi.org/10.1007/978-94-017-1618-5

Lipton, Peter. 1991. *Inference to the Best Explanation*. London: Routledge.

Lipton, Peter. 1993. “Is the Best Good Enough?” *Proceedings of the Aristotelian Society*, 93: 89–104. https://doi.org/10.1093/aris

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

Musgrave, Alan. 1988. The Ultimate Argument for Scientific Realism. In: *Relativism and Realism in Science*, edited by R. Nola, 229–252. Dordrecht: Kluwer Academic. https://doi.org/10.1007/978-94-009-2877-0_10

Roche, William and Sober, Elliott. 2013. “Explanatoriness Is Evidentially Irrelevant, or Inference to the Best Explanation meets Bayesian Confirmation Theory.” *Analysis*, 73: 659–68. https://doi.org/10.1093/analys/ant079

van Fraassen, Bastiaan C. 1989. *Laws and Symmetry*. Oxford: Clarendon Press. https://doi.org/10.1093/0198248601.001.0001