BALKAN JOURNAL OF PHILOSOPHY

Vol. 18, Issue 2, 2026

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

TABLE OF CONTENTS

Invited Paper

Truth and Modality (III) / 187

Fabrice Pataut

The Handicap Principle and Norms of Assertion: Why Social Norms Cannot Function as Stable Cost-Imposing Mechanisms / 197

Dimitar Elchinov

Embodiment, Playfulness, and Humanization / 211

David-Augustin Mândruț

How Are Large Language Models Fallible: Epistemic Prospects for Neural Intelligence / 223

Boris Grozdanoff

Human Transformation in the Digital Age / 239

Viktor I. Mishchenko, Larysa M. Gren, Olga V. Shimchenko

The Roots of Practical Realism and Metaphysics / 254

Zhumagul Bekenova, Peeter Müürsepp

The Influence of 'Philosophy of the Illuminated' in the Khomeini's Approach to Create an 'Ideal Society' / 271

Bogdana Todorova

Personhood in African Environmental Ethics / 283

Terzungwe Inja, Rotimi Omosulu

Main Features of Existentialism in French Literature (Based on the Works of J.M.G. Le Clézio and P. Modiano) / 293

Rasima Asadzade

The Fundamental Principles, Arguments, and Implications of Existentialism in the Light of Sartre / 312

Gizem Şerifoğulları, Abdülkadir Çüçen

Self-Realization as a Way to Freedom: Plato and Spinoza in Dialogue / 328

Junshuo Bian

Short communication

Reassessing Plethon's Critique of Aristotle: Responses from Trapezuntios and Bessarion / 346

Márton Adorján Kecskés

Book reviews

Milkov, N. (2026). *The Tractatus, Wittgenstein's Method and Analytic Philosophy* / 351

Kamen Lozev

Ioannidis, S., & Psillos, S. (2022). *Mechanisms in Science: Method or Metaphysics?* / 357

Aida Roige, Saúl Pérez-González

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 187 – 196

Invited paper

TRUTH AND MODALITIES (III)[1]

Fabrice Pataut*

Abstract: I contrast two construals of the thesis that truth is independent of verifiability in principle: a modal one and a non modal one. I argue in favor of the modal construal and then, on that basis, that independence holds across the board, i.e., even for statements that are verifiable by us relative to familiar, customary, non-skeptical standards.

Keywords: contingency, disquotation, epistemic possibility, truth, Dummett (Michael A. E., Sir), Wright (Crispin).

1. Realism and independence

2. Possibilities

3. Contingency

4. Conclusions

References

3. Contingency

Epistemic possibility is to be understood here in terms of a relation between what is known and what is possible: it is epistemically possible that s at a world W just in case the possibility that s (or that s be true) is either consistent with what is known at W or follows from it. Whether or not it is known by anyone that s (or that s is true) at W is a contingent matter in the following sense: whoever happens to have that knowledge could very well not have it. Knowledge of the truth of s, or of its compatibility with any other statement, or of what follows from it at W, logically or otherwise, is also a contingent matter. The notion of knowledge we need in order to unpack the consequent in

(2) $\Box$ (s is true $\rightarrow$ s is verifiable by us)

so that, once (2) is argued for, the opposite claim that

(1) $\Diamond$ (s is true $\cdot$ s is unverifiable by us)

[1]The first and second parts of the article were published in *Balkan Journal of Philosophy* 14(2), (2022) 103–108 and 15(2), (2023) 95–100, respectively. Numbered sentences have kept the numbers they were assigned in parts I and II of the article.

*CNRS, UMR 8011 Sciences, Normes, Démocratie Paris, France; patautfabrice@gmail.com

©2026 Fabrice Pataut. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

is refuted, quite naturally turns all truths into truths that are knowable at least in principle, although not necessarily into contingent ones. By parity, the notion of knowledge we need in order to unpack the second conjunct in (1) so that (2) is refuted turns all truths into possibly unknown ones, although not necessarily into contingent ones. In particular, in either case, mathematical truths, logical truths and metalogical truths (the truths of the theory of sentences or formulae used to express logical truths) may still be deemed necessary.

In other words, what turns out to be contingent isn't the truth per se of statements, but either the fact that our knowledge of them must be acquired, at least in principle (in case (2) is vindicated), or the fact that the very same knowledge could very well not be acquired (in case (1) turns out to be vindicated). What is known at a world W depends on which evidence in favor of s knowers are able to uncover at W. The fact that the truth of s is either known or ignored turns out to be an entirely contingent matter even when s is deemed necessarily true (if true at all).

It has also been remarked that the relation between epistemic and conceptual possibilities is such that the truth of the consequent in

(14) $\Diamond_{EPIST}$ (s is true · s is unverifiable by us) $\rightarrow$ $\Diamond_{CONCEPT}$ (s is true · s is unverifiable by us)

qua conceptual would remain intrinsically beyond our ken even if it could be argued for on the basis of an argument from contingency (Pataut, 2023, p. 99). But it still remains to be determined whether an argument may be proposed against the view that s being possibly true independently of our being able to recognize that it is should be thought of as something we might at best *conceive* but not as something we may actually *know*.

I shall now move on to the argument in favor of the view that this is indeed something we know, as opposed to something we may conceive without any assurance that the concept thus formed has any ground or legitimacy. Contingency will be a key element in the argument.

I shall express contingency in the following way: ∇ will be the monadic statement-forming operator for contingency, so that: $\nabla s =_{Def.} s \cdot \Diamond \neg s$ (see, e.g., Montgomery & Routley, 1966).

What we need now is a reformulation of realism in terms of possibly unknown truths. In the following reformulation of (1):

(15) $\Diamond$ (s is true · $\neg Ks$)

the knowledge operator K will be read as "It is known that."

Given that $\nabla Ks =_{Def.} Ks \cdot \Diamond \neg Ks$, the argument from contingency must allow us to conclude that:

(16) $\nabla(Ks)$

given that we could reformulate (16) as:

(17) ∇(*Ks* is true). Given that $\nabla Ks =_{Def.} Ks \cdot \Diamond \neg Ks$, the argument from contingency must allow us to conclude that:

(16) $\nabla(Ks)$

given that we could reformulate (16) as:

(17) ∇(*Ks* is true).

Since we need to rule out the idea that truth should be constrained by knowledge (just as we must rule out the idea that it should be constrained by verifiability in principle or by verification hic et nunc, either being a reliable source of knowledge), we must give a philosophical content to the formal notions of possibility, contingency and knowledge that are available to the realist in his attack on the strong modal claim (2) and in the defense of his own position. With (1), (15), (16) and (17), we've only provided formal claims expressing the position to be defended. I shall argue that the adequate notion of possibility for the defense of realism is the epistemic notion. I shall first defend the view negatively against the argument to the effect that because assertions and denials of truth are normative in character, the debate commits us to a substantive notion of truth, i.e., to a notion over and above the disquotational "true" (Wright, 1995)[2]. I shall then turn to the notion of contingency we need in order to account for the possibility that evidential relations may break down, thereby allowing us to conclude that either (1) or (15) holds, given that our access to whatever allows us to determine that *s*, or that *s* is true is neither necessary nor impossible.

To begin with, should anyone defending (1) or (15) be committed to a substantive notion of truth, i.e., to a notion over and above the disquotational "true" which would involve the endorsement of objectively determinate semantic relations (perhaps so-called "correspondence" relations) between the statements of the language for which we have defined or constructed a truth predicate, and truth-conferring facts or states of affairs we might entirely (or partially) ignore? If this were the case, the defender would have to argue that either (1) or (15) is to be accepted because the relevant semantic relations between *s* and "the world," or whatever makes it the case that *s* is true, are not always, or necessarily, available to us. The reason might be that, for all we know, evidential semantic relations may and do sometimes, break down.

Wright, for instance, claims that the deflationary conception of truth is ill-equipped for the expression of normative claims – either to affirm or to deny them. Since assertions and denials of truth are both normative, the deflationist

> *cannot* allow that [...] "true," when used to endorse, has the function of commending a proposition for its satisfaction of *some distinctive norm* which contrasts with epistemic justification and which only "true" and equivalents serve

[2]Williamson's contention that the antirealist needs a stronger *non*-epistemic or so-called "metaphysical" notion is also relevant here (see Williamson, 2000, p. 273). If Williamson's claim is correct, the realist needs an even stronger non-epistemic or so-called metaphysical notion of truth so that realist truth, qua truth beyond all possible verification or verifiability, jettisons epistemic considerations pertaining to belief and knowledge for reasons that should be *incompatible* with the reasons that lead his opponent to do the same. It should be clear why the antirealist's metaphysical commitments are too weak. I shall not comment on this question here.

> to mark. For if there were a distinctive such norm, it could hardly fail to be reckoned a genuine *property* [my emphasis] of a proposition that it did, or did not comply with it. And if the norm in question were uniquely associated with "true" and its cognates, that would be as much as to allow that there was a special property of *truth* – at which point the deflationary game would have to be given away. (Wright, 1995, p. 212)

Wright's objection is that the deflationist has insuperable difficulties with objective norms because our recognition of norms and our commitment to them involves the endorsement of a property. Wright thinks that this is so because

> it's intuitively perfectly evident that the use of "true" *is* tied to a norm – to a way in which acceptance of a proposition may be in good, or bad standing – quite *separate* [my emphasis] from the question of its justification in the light of the acceptor's state of information. (Wright, 1995, p. 214)

But the distinction on which Wright relies here to sustain his claim begs the question. We can't possibly *rely* on the idea that, say, "snow is white" is a candidate for denial if *we* have no warrant for it, although it might nevertheless be in bad standing in quite another way, namely if *it is not the case* that snow is white *whether or not there is any available warrant for it*. Two points must be made in this respect. To begin with, we can't consider it established that "the concept of truth is a concept of a way a proposition may or may not be in good standing which precisely *contrasts* with its justificatory status at any particular time" (Wright, 1995, p. 214). What we need is an argument for it. Secondly, in order to claim that the proposition expressed by "snow is white" (if any) satisfies a norm distinct from epistemic justification, we need not go so far as to claim that "snow is white," although true and detectably true by us, might nevertheless, for all we know, be false.

Another point made by Wright is that because deflationary truth is a device allowing us to generalize in a way that can't be done in the object language, we, as it were, restore the effect of objective reference to non-linguistic items (to "reality") by asserting something in the object language. If the idea is that we must descend from semantic ascent in order to express the contentious realist claim, then the criticism is misguided. Someone might perfectly well claim that disquotational truth could transcend provability or verifiability, either *hic et nunc* or in principle. Why should the claim that the truth predicate is a device for generalizing assertions make that impossible?

For any statement s of a language L, no more is needed in terms of truth than the metalinguistic variant of *s* ("*s*" is true) in order for (17) to provide a ground for (15). One might for instance want to assert that it could be the case that for every even number z greater than 2, there are prime numbers x, y, such that $z = x + y$ although we are unable to either prove or disprove the conjecture. Or one may want to assert that it could be the case that Caesar crossed the Rubicon although we are not in a position to verify or falsify the claim. The disquotational "true" will be his or her device for generalizing assertions belonging to a given domain, say mathematics or history, i.e., assertions that may be expressed, respectively, in the language of mathematics, or in that part of natural language containing predicates and singular terms suited to

the appropriate expression of historical claims. He or she will be asserting: "For all *s* (of a given class), it could be the case that *s* were true even though we are not in a position to determine that it is."[3]

We may want to consider two distinct cases in this respect: cases in which the language does not contain its own truth-predicate, and cases in which it does, cases in which "True-in-*L*" may not be represented in *L*, and cases in which it may.

In the first Tarski-type case, L_1 contains a truth-predicate (or, better: a satisfaction predicate) T_1 for L_0, so that when we iterate the process, we have a sequence of languages $\{L_0, L_1, L_2, L_3, \dots\}$, each with a truth predicate $T_1, T_2, T_3, \dots$ for the preceding language in the series.

In the second case, we might, e.g., in the manner of Martin and Woodruff (Martin and Woodruff, 1975) or, better, in the manner of Kripke (Kripke, 1975), consider that the extension and counterextension of the truth-predicate for a language *L* are both empty, so that both attributions and denials of truth are themselves without truth-value. The sentences of *L* which do *not* contain occurrences of the truth-predicate are semantically evaluated intuitively. Some turn out to be true while others turn out to be false. These two classes of evaluated sentences constitute, respectively, the extension and counterextension of the truth-predicate at the next level up, so that more sentences are evaluated in semantic terms of truth and falsity at this level. The process of chain valuation may be continued or iterated into the transfinite. A truth-representation theorem for *L* is then proved showing that the resulting extensions of the truth-predicate all collapse into a "fixed point," i.e., a point at which all sentences which will belong to the extension of the predicate in the chain process *already* do.

Now, perhaps, as Parsons has suggested, Tarski's

> hierarchical approach has seemed implausible in application to natural language because there seemed to be no division of natural language into a hierarchy of 'languages' so that the higher ones contain the 'semantics' of the lower ones. (Parsons, 1974, p. 382)

In any case, the point about the disquotational or deflationary "true," or use of "true," being sufficient for a disagreement of the kind we are considering to emerge, should be independent of such considerations and hold no matter how truth-in-*L* might end up being represented, either infralinguistically or meta linguistically.

The answer to the objection that any notion of truth according to which truth may transcend verifiability or provability by us, either hic et nunc, in the long run of scientific inquiry, or at any point in time (in an indefinitely extensible past, or in an indefinitely extensible future) *is* a substantive notion that the deflationist fails to capture, amounts to this: let the disquotational notion of truth *be* that substantive notion in the following sense, or with the following proviso. Let the particular instances of Tarski's T-schema for a language *L* be sufficient for the advocate of (1) or (15) to express his

[3] As a matter of fact, the point I'm making here is only accidentally connected with generalization as such. Even if, *per impossibile*, the truth predicate didn't have that function, it remains that no realist in Dummett's sense could allow either (1) or (15), in particular if they were to be argued for on the basis of (16) or (17).

or her disagreement about truth. In other words, let us take truth-conditions to be redundant truth-conditions, so that "s is true" states no more than what s alone already states *without* the truth-predicate attached to it.

The point here is that the substance, thickness or import of the notion, consisting in its independence from verifiability or provability, is adequately captured by the metalinguistic (or infralinguistic) variant of each *s* of *L* for any given class of statements. It *need not* be captured by alleged objective semantic correspondences which could obtain without us knowing that they do. In any event, no antirealist of the Dummettian variety will allow, for disquotational "true," that it is possible for *s* to be true and yet be unverifiable by us. When attacking the modal claim, Dummett's antirealist need not target objective semantic relations between expressions and their referents, say between singular terms and objects, predicates and extensions, statements and states of affairs. He may take "*s* is true" in (1) to be nothing more than a metalinguistic or infralinguistic variant of *s*, stating, either in the metalanguage in which the expression "is true" is available, or in the object-language itself, the contentious claim he wishes to refute.

By the same token, the antirealist might very well *not* be satisfied with the pragmatist position defended by Peirce and James, according to which statements are true insofar as they are useful, save labor, make a difference in one's actual life and, in general, just insofar as the truths they happen to express depend on our practical interests (see James, 1907 and Peirce, [1905] 1998). If it is also claimed that we could ignore the usefulness of truth, or be mistaken about our own practical interests, the pragmatic considerations constraining truth might very well turn out to be as inaccessible as any transcendent notion the pragmatists originally wished to get rid of (say, as an abstraction without psychological dimension).

In other words, the conclusion to be drawn here is that the antirealist must make sure that whatever conditions will turn out to constrain truth-conditions should be such that it must be possible, at least in principle, perhaps even effectively, to acknowledge that they obtain when they do.

Let me now turn to the philosophical content of the notion of contingency that is needed for our argument in favor of the opposite realist claim. Contingent states of affairs are neither necessary nor impossible. If the obtaining of knowledge is contingent in this sense, the modal notion in (1) *must* be epistemological. The contingency of the connection between truth-bearers and our recognitional capacities with respect to truth or truth-conditions in the disquotational sense *provides* the adequate ground for possibility of the epistemic kind.

We had said earlier (Pataut, 2023, p. 98) that

(9*) The possibility in (1) is epistemic if and only if (1) follows from what we know.

We may now modify (9*) so that:

(9**) The possibility in (1) is epistemic if and only if (1) follows from either (16) or (17).

Given that both (16) and (17) are indeed known and not merely conceived – or given that our conception of (16) and (17) being the case is grounded in the way the relevant recognitional capacities are to be exercised – then both (1) and (15) hold.

4. Conclusions

The fundamental point of antirealism construed in terms of knowledge and possible worlds is that if s is true at a world, then, at least in principle, s is knowable at that very same world. Two notions are at stake in (1): the semantic notion of truth and the modal notion of possibility. Dummett's so-called "manifestability" argument directly attacks the possibility of verification-transcendent truth (see, e.g., Dummett, [1973] 1978)[4]. The argument from contingency insists that the possibility of the loss or breaking down of the evidential relations that would justify knowledge claims at a given world provides grounds for rejecting the fundamental point. This also holds with respect to accessibility relations. If we understand the expression "would have been known at W_n" as "is known in a world accessible from W_{n-1}" and if verification and proof procedures available in W_{n-1} determine which worlds are accessible from it, then the argument from contingency must also apply, at least in the sense that accessibility relations might also be ignored.

The idea that has been defended throughout is that truth is contingently connected to what we are able to achieve in terms of proofs and the acquisition of empirical warrants. It is indeed granted that as-yet undiscovered proofs and inaccessible empirical evidence are no less transcendent and thereby antirealistically suspect than inaccessible truth. It would indeed be a mistake to assume that what justifies our knowledge claims at a given world is, as it were, waiting for our cognitive abilities, skills and expertise pertaining to a given domain to be activated.

With respect to the mathematical case, this would be tantamount to assuming that, say, there is a proof of Goldbach's conjecture that we have failed to discover so that Goldbach's conjecture might be true although we are not yet able to prove it: unbeknownst to us it might be that for all even integers z strictly greater than 2, there is a proof to the effect that there are prime numbers x, y, such that $z = x + y$. Or this would be tantamount to assuming that, say, Fermat's last theorem could be true independently of the reason we now have at our disposal to believe that it is, i.e., Andrew Wiles's proof of 1995. We know at our world that there are no positive integers x, y, z, n with n ¿ 2 such that $x^n + y^n = z^n$. We previously had partial proofs restricted in scope, at other worlds, for exponents 3 (Euler), 4 (Fermat and Euler), 5 (Dirichlet and Legendre), 6 and 7 (Lamé and Lebesgue) so that, at different worlds, we had different grounds for different knowledge claims. But assuming the existence of a yet unknown general proof such as Wiles's, or of any of these partial proofs is clearly a mistake.

[4] It attacks the semantic principle of bivalence and the truth-conditional theory of meaning somewhat less directly.

The crucial point when addressing this legitimate worry about the contingency argument is that no platonistically construed proofs or verifiability-transcendent empirical data have been smuggled into the appeal to the fortuitous or accidental character of the recovery or construction of warrants.

Now the antirealist might still be worried that (2) hasn't been properly refuted. He or she may now object that a proper rebuttal of (2) *must* be grounded in an appeal to some specific recognitional capacity that we should be able to activate with respect to (1) or (15). How legitimate is that worry?

I'm not claiming here that it is illegitimate, or even wrongheaded, to link transcendence to practical recognitional capacities as such. I'm claiming that there is no reason why we should look for a recognitional capacity with respect to (1) or (15) that must be modelled on the one that, once properly activated, satisfies the antirealist with respect to "*s* is true."

The point Dummett's antirealist is making is in part epistemological: it must follow from conditions imposed on our linguistic practice that we can't manifest a knowledge of transcendent truth-conditions. The argument is best conceived of as a challenge rather than as a *reductio ad absurdum* of the modal thesis, a challenge yielding a provisionally negative conclusion to the effect that, until we can make it manifest, by the exercise of appropriate recognitional capacities, that truth is independent of the possibility of its recognition by us, we should conclude that the only *bona fide* notion of truth is one constrained by the epistemic notion fitted to particular areas of discourse: gradual, defeasible and partial verifiability in the empirical case, formal provability or syntactic derivability in the logico-mathematical case[5].

It isn't simply, of course, that this happens to be the only notion of truth we may manifest in linguistic practice, make legitimate use of when carrying out deductions, or truly master, as if there still could be some *other* legitimate notion out there, we have no access to. Both manifestability in principle and effective or actual manifestation are supposed to pay dividends. The point of the negative argument is that if we are unable to manifest a grasp of any other notion of truth, or can't do it in a finite time, then this must be what truth amounts to. There is no notion of truth over and above the manifestability requirement; the possibility of there being some transcendent or inaccessible notion is thereby nipped in the bud.

What I have suggested here is that it is possible to defend (1) or (15) without the help of a recognitional ability modelled on the one we activate when determining the truth of *s*.

At this point, the advocate of (2) might object that classes of decidable statements and classes of *pro tempora* undecided statements cannot be treated on a par. The objection would be that the lack of discrimination between the two kinds of cases would force us to conclude that, whenever we are in a privileged epistemic position with respect to statements of a given class, i.e., when we are able to determine that they

[5]This construal is preferable to the view according to which Dummett's antirealist eschews the notion of truth altogether, strictly speaking *substituting* verifiability and provability for truth, and verifiability-conditions and provability-conditions to truth-conditions.

are true, their truth could still be determined *differently than we justifiably believe it is*. In other words, there might be warrants or kinds of warrants unknown to us. If a statement is independently true even when we are in a position to determine that it is indeed true, i.e., even when we have as a matter of fact implemented a process of truth determination and have come to its completion in a finite time, then it might very well have entirely different truth-makers.

This amounts to a skeptical position with respect to the statements of the given class, akin to the kind of position Descartes sought to refute when considering the evil genius hypothesis, or that Putnam judged self-refuting when considering the hypothesis that we might be brains in a vat (Putnam, 1981, ch. 1).

The problem, then, is to determine whether or not it is consistent to defend the weaker independence claim, restricted to classes of *pro tempora* undecided statements, without also (or thereby) defending the stronger one, which holds across the board, i. e., for all classes of statements, including the decidable ones. In particular, is a realist committed to skepticism, or to what Putnam once called "metaphysical realism," or to anything which would lead to it, no matter how indirectly?

The correct answer to these worries, I think, is that although it is consistent to defend the weak claim without thereby defending the strong one, one may endorse the modal claim which has often been taken to be the earmark of realism in Dummett's sense, *without* endorsing metaphysical realism, or being led to a skeptical dead end. To vindicate realism thus conceived, one needs to find an argument showing that the independence of truth from its recognition *by us* follows from a conception of the contingencies on which our acquisition of knowledge depends. This is where the epistemic notion of possibility comes in. The point isn't merely that the possibility that s be true and not provable or verifiable by us is syntactically consistent with such a conception, but that such a conception of contingency *entails* the disputed modal claim.

I have suggested that the disagreement lies precisely in the notion of possibility that a realist may rely on in order to defend his modal thesis, and which the antirealist must reject. When classes of decidable statements (and of effectively decided ones) are taken into consideration, there remains a more fundamental ground for the contention that these statements could be true beyond all possible proofs or verifications, not because some fancy evil genius is after us, or because of some highly improbable brains-in-a-vat situation, but, once again, relative to normal *unskeptical* standards.

The correct formulation of the debate must take into account statements, which are recognized to be true by unskeptical standards. The point is *not* just epistemological: although these statements are conclusively proved or verified, there is a more fundamental question of whether they could *logically*, not *epistemologically* be true beyond all possible proofs or verifications.

In other words, the weak claim divorces truth only once: from our abilities to acknowledge truth, but *not* from the evidence there is in its favor whenever such evidence happens to be available. The strong claim divorces truth twice: both from the abilities we have at our disposal to acknowledge it *and* from the evidence itself

on which such an acknowledgement is to be grounded. If the possibility of error remains when we have been successful in our research, the stronger claim goes as far as divorcing truth from (alleged or illusory) recovered evidence or actual proofs. We certainly don't need to go that far.

References

Dummett, M. ([1973] 1978). The Philosophical Basis of Intuitionistic Logic. *Truth and other Enigmas*, 1st ed., Cambridge, MA: Harvard UP, 215–247.

James, W. (1907). Pragmatism's Conception of Truth. *The Journal of Philosophy, Psychology and Scientific Methods* 4(6) [March 14 1907], 141–155.

Kripke, S. (1975). Outline of a Theory of Truth. *The Journal of Philosophy* 72 (19), 690–716.

Martin, R. & Woodruff, P. (1975). On Representing 'True-in-L'. *Philosophia (Ramat-Gan)* 5(3), 213–217.

Montgomery, H. & Routley, R. (1966). Contingency and non-contingency bases for normal modal logics. *Logique et Analyse* 9(35–36), 318–328.

Parsons, C. (1974). The Liar paradox. *Journal of Philosophical Logic* 3(4), 381–412.

Pataut, F. (2023). Truth and Modalities (II). *Balkan Journal of Philosophy* 15(2), 95–100.

Peirce, C. S. ([1905] 1998). What Pragmatism Is. *The Essential Peirce*, vol. 2, The Peirce Edition Project. Bloomington: Indiana UP, 331–345.

Putnam, H. (1981). *Reason, Truth and History*, Cambridge MA: Cambridge UP.

Williamson, T. (2000). *Knowledge and its Limits*, Oxford: Oxford UP.

Wright, C. (1998). Truth: A Traditional Debate Reviewed. *Canadian Journal of Philosophy*, Supplementary Volume (24: Pragmatism), 31–74.

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 197 – 210

THE HANDICAP PRINCIPLE AND NORMS OF ASSERTION: WHY SOCIAL NORMS CANNOT FUNCTION AS STABLE COST-IMPOSING MECHANISMS

Dimitar Elchinov*

Abstract: This paper examines philosophical applications of Zahavi's handicap principle to assertion reliability, focusing on contributions by M. Green, P. J. Graham, and J. P. Bruner. The handicap principle holds that honest signals are costly to produce and that this cost is differential – it falls more heavily on lower-quality signallers than on higher-quality ones, which is what prevents deception and keeps signals reliable. I argue that social norms cannot function as cost-imposing mechanisms in handicap-based accounts of assertion.

My argument consists of two stages. First, I use Penn and Számadó's biological objections to preemptively undermine any naturalistic interpretation of the way in which the handicap principle has been deployed in philosophy, and, in particular, to block any such naturalistic interpretation of the concept of "social norms." Second, I examine the notion of "social norms" as used by philosophers to argue that biological handicaps impose costs deterministically, whereas social norms impose costs only probabilistically, and that the social costs of assertion introduce systematic variability that renders social norms incapable of serving as stable cost-imposing mechanisms.

Although I do not offer a positive account of the philosophical application of the handicap principle, I think that my contribution may nevertheless be helpful insofar as it identifies serious obstacles to such deployment.

Keywords: handicap principle, norms of assertion, social norms, costly signaling, biological metaphors, speech act theory.

1. Introduction

The reliability of human communication poses the "assertoric puzzle": how do assertions succeed in conveying accurate information when speakers have obvious incentives to deceive? Mitchell Green (2009) proposed that norms of assertion function as handicaps ensuring reliability through cost imposition. Peter Graham (2020) argued for deterrent rather than handicap mechanisms, while Justin Bruner (2024) introduced internal sanctions as handicap-like mechanisms that overcome the limitations of both external deterrents and classical handicaps. This philosophical debate unfolds against significant developments in ethology itself, where Penn and Számadó (2020) argued that the handicap principle, although influential, is fundamentally flawed as a general explanation for honest signaling. It is important to note that this paper takes no position on whether human communication is reliable; that question belongs to the accounts under scrutiny, not to the scrutinizer.

*Sofia University St. Kliment Ohridski, Bulgaria; dgivanov@phls.uni-sofia.bg

©2026 Dimitar Elchinov. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

Although philosophical contributions have, up to now, not considered the scientific problems of the handicap principle, there is no reason to remove it from philosophical debate, even if the principle itself is nothing more than a product of imagination. However, I claim that the biological issues associated with the handicap principle complicate efforts to use it for building accounts of norms of assertion and that, notwithstanding these issues, philosophers have failed to make sense of the application of the handicap principle because all attempts to do so involve reference to social norms. I argue that social norms, if "norms" at all, cannot impose costs on assertions, due to the way they are conceptualized in research literature. To justify my claims, I will first review relevant biological contributions and show how the possible rejection of the handicap principle in biology impedes the naturalistic character of philosophical contribution. Second, I will summarize the ongoing debate on the application of the handicap principle in philosophy. Third, I will show how philosophical efforts at establishing an analogy between handicaps and "costly" assertions fail because of their inability to match the deterministic manner in which costs are imposed in biology with the probabilistic way in which costs "may" be imposed through social norms.

2. The biological foundation: Zahavi's handicap principle

Amotz Zahavi's handicap principle, proposed in 1975, attempts to explain why costly secondary sexual characteristics that apparently reduce survival fitness nonetheless persist through natural selection. Zahavi originally formulated the principle by stating that "an individual with a well-developed sexually selected character is an individual which has survived a test" and that "females which selected males with the most developed characters can be sure that they have selected from among the best genotypes of the male population" (Zahavi, 1975, p. 207). The classic example, the peacock's elaborate tail, makes the bird conspicuous to predators and hampers flight, yet persists because it functions precisely through its costliness. That is, for signals to be honest, they must be costly to produce, with costs that are differential (more expensive for lower-quality signalers than higher-quality ones). This cost structure prevents deception because inferior individuals cannot afford false signals of high quality. The peacock's tail serves as an honest advertisement of genetic fitness precisely because maintaining such elaborate ornamentation requires genuine biological resources.

Alan Grafen's formalization provided a mathematical presentation for Zahavi's intuitions, demonstrating how costly signaling could maintain honesty in evolutionary equilibria. This mathematical model demonstrates signaling equilibria where individuals make decisions based on their quality and differential cost structures. According to Grafen's formal model, "it costs a better male less to make the same signal" (Grafen, 1990, p. 532). This condition is key to separating equilibria between signalers of different quality levels. The mathematical framework ensures that each male type chooses the level of advertising that maximizes his expected reproductive success (ibid., p. 521). Grafen interpreted his model as biological validation of Zahavi's handicap principle, explicitly stating that "Zahavi's major claims for the handicap principle are thus vindicated" (ibid., p. 517) and that "the evolutionary stability of persuasive signalling necessitates honesty, which necessitates waste" (Grafen, 1990, p. 532).

However, according to Penn and Számadó (2020), this interpretation was mistaken as Grafen's models actually demonstrate condition-dependent signaling where individuals optimize investment based on their quality rather than engaging in wasteful overinvestment. This aligns with Darwinian principles where natural selection typically eliminates inefficiency and waste in favor of optimal resource allocation: "Rather than being wasteful over-investments, honest signals evolve in this scenario because selection favours efficient and optimal investment into signal expression and minimizes signalling costs" (Penn and Számadó, 2020, p. 267).

Penn and Számadó's comprehensive argument challenges the theoretical foundations entirely: "[t]he most widely cited explanation for the evolution of reliable signals is Zahavi's so-called Handicap Principle, which proposes that signals are honest because they are costly to produce" but "this idea is erroneous" and became accepted through "confirmation bias" and "inadequate testing" (ibid.). Their attack against Zahavi's formulation of the handicap principle is based on several points: many honest signals in nature are not costly at equilibrium (contradicting the principle's core prediction); Grafen's models support condition-dependent signaling rather than wasteful signaling; and theoretical work shows honesty can be maintained through differential benefits rather than differential costs. As they conclude, "signalling costs paid at the equilibrium are neither sufficient nor necessary to maintain signal honesty, and honesty can evolve through differential benefits, as well as differential costs" (Penn and Számadó, 2020, p. 283).

The theoretical problems identified by Penn and Számadó are complemented by empirical disconfirmation of the handicap principle's core predictions. "Many studies refute the assumptions that signals are usually honest [...] and very costly to produce [...]. [...] [M]oreover there is no evidence that signals are honest because they are costly" (Penn and Számadó, 2020, p. 269–270). According to them, the empirical evidence systematically contradicts handicap mechanisms: "The number of empirical studies that contradict the assumptions and predictions of the Handicap Principle are increasing steadily" and "These anomalies have largely been ignored and theoreticians have offered no way to resolve these problems" (Penn and Számadó, 2020, p. 270). This pattern reflects what they identify as widespread "confirmatory bias" in both theoretical and empirical research (p. 286).

3. Why is this a problem for philosophy?

These biological objections do not necessarily render all applications of reasoning involving the handicap principle in philosophy implausible. Moreover, metaphors, such as the metaphor of the handicap principle in the philosophical debate on norms of assertion, can function well enough even when their source domain contains false statements. However, Penn and Számadó's objections suggest that careful consideration is required when and how such reasoning applies, particularly given the ongoing debates in biology about signaling mechanisms. The philosophical application of the handicap principle is treated as a metaphor: not that human communication involves anything biologically analogous to peacock tails or metabolic costs, but that the principle's structural logic, that differential costs can stabilize honest signalling, is borrowed

as a template and applied to a different domain. Its explanatory success in that domain depends on whether the structural role can be filled. The question is whether anything in the domain of human assertion plays the same role that differential metabolic cost plays in biology.

In case the objections of Penn and Számadó overturn the perceived acceptance of the handicap principle, the models offered by philosophers and focusing on "costly" assertions as a correlate to the handicap principle in biology lose scientific grounding and are rendered as more or less successful captures of the everyday intuition that "actions have consequences". From a metatheoretical perspective, this would pose an obstacle to accounts that are contingent on the scientific status of the handicap principle to claim that the premises and the concepts they deploy are consistent with naturalism. In particular, if one could simply draw on an analogy with cost-imposing mechanisms in ethology and stipulate that "social norms" function as natural, deterministic cost-imposing mechanisms, bypassing the variability problems documented in social science research, they might rescue handicap-based accounts of assertion from my objection. However, the scientific doubt cast on the handicap principle in its original domain blocks this theoretical escape route. Since the biological foundation itself is questionable, one cannot appeal to consensus in biology to justify treating social norms as natural kinds with law-like cost-imposing properties. This forces philosophical applications back onto the standard understanding of norms in the social sciences as inherently variable phenomena, which, as I will demonstrate, cannot provide the stable cost structures required by handicap mechanisms.

4. Green's philosophical application

Green (2009) recognized the handicap principle's potential relevance for a central problem in speech act theory: how assertions reliably express beliefs (Green, 2009, p. 141). According to expressivism, asserting *p* expresses the belief that *p*, but this raises the question of whether and how assertions provide strong evidence for speakers' psychological states when speakers have obvious incentives to deceive.

Green proposed that norms of assertion function as handicapping mechanisms ensuring reliable expression of belief. On this account, norms governing assertion are analogous to handicaps in the biological sense: they impose costs that help to ensure the reliability of the signals. Just as the peacock's tail serves as an honest signal of genetic fitness through costliness, norms of assertion might serve as honest signals of belief through social costs imposed on violators. When speakers assert *p*, they subject themselves to social costs if *p* turns out false (loss of credibility, social disapproval, potential exclusion from future epistemic exchanges) (Green, 2009, p. 153). These costs function as handicaps because they make false assertion expensive, particularly for speakers lacking good evidence for their claims. Yet to function, these cost-imposing mechanisms involve testing the agent's expressions against social norms:

> A norm can make a signal costly to produce not by exacting calories or territory, but rather by making an agent subject to a loss of credibility. For a norm might

> have the following schematic content: One who produces S is to be in condition C; otherwise she is subject to a loss of credibility.
>
> A loss of credibility will, as we have suggested, reduce the weight of a speaker's speech acts, thereby hobbling her ability to serve as a conversational "player". As the likelihood and severity of a loss of credibility increase, so does the exposure to risk for those who produce S without being in condition C. (Green, 2009, p. 154)

The differential cost imposition emerges because speakers with good evidence for p face lower expected costs than speakers without such evidence. A speaker with strong evidence expects their assertion to be true and faces minimal risk of social sanctions. A speaker without good evidence faces a substantial risk of being revealed as unreliable. This would encourage honest assertion and discourage deceptive or careless claims. Green argued that this liability-based mechanism explains the stability of assertoric communication by securing the reliability of assertive signals (Green, 2009, p. 152), clarifies why hearers typically trust assertions absent defeaters by making assertion provide strong evidence of sincerity and justification (p. 157), and illuminates the social grounding of epistemic normativity, since credibility and epistemic authority are governed by socially enforced norms (p. 153).

5. Graham's deterrent alternative

Graham (2020) mounted a systematic critique of Green's handicap model, arguing for significant dissimilarities between biological handicaps and social norms governing assertion. Graham argues that in "traditional" handicaps like the peacock's tail, costs are paid at equilibrium by honest signallers (the peacock continuously bears metabolic costs of maintaining elaborate plumage). In assertion, however, costs are only paid off equilibrium by dishonest speakers caught in false assertions. As Graham explains, "When a quality-handicap stabilizes honesty, costs are paid within the stable strategy; the costs are actually paid by honest signalers within equilibrium. When a deterrent stabilizes honesty, on the other hand, costs are paid by the dishonest signaler who strays from the stable strategy, outside of equilibrium" (Graham, 2020, p. 9).

Graham also argues that biological handicaps reliably signal only the quality they waste (the peacock's tail signals genetic fitness precisely because it wastes metabolic resources correlating with fitness). But social costs of false assertion don't obviously waste any quality correlating with epistemic reliability. Graham argues that there's a good reason assertions and other speech acts in general can't be quality-handicaps, for "quality-handicaps can only reliably (stably) signal the quality they waste; that's how the mechanism works" (ibid., p. 11), but "when we use language, we can signal – we can mean – just about anything. Zahavi's mechanism can't be the mechanism that explains the stability of human communication, at least not in general" (ibid.).

Graham's deterrent model suggests that costs are imposed externally by communities rather than being intrinsic features of signaling mechanisms. On this model, deterrent mechanisms maintain honesty not because assertion is intrinsically costly, but because communities make it costly to be caught in dishonest assertion.

6. Bruner's Internal Sanctions Approach

Bruner's (2024) analysis acknowledges weaknesses in Graham's deterrent model while proposing a novel solution that rehabilitates handicap-like mechanisms. His central insight is that external deterrents often fail to provide strong evidence for speaker belief, but internal sanctions (guilt, shame) can function as handicap-like mechanisms that overcome the limitations of both external deterrents and classical handicaps. Bruner's main contribution is the distinction between receiver-dependent costs (deterrents imposed by listeners) and receiver-independent costs (intrinsic to the signalling act itself).

The problem with Graham's deterrent model, according to Bruner, is that external deterrents often fail to provide strong evidence for speaker belief (Bruner, 2024, p. 3). In deterrent-based systems, costs are receiver-dependent and imposed externally by communities regardless of speakers' underlying evidential positions. Speakers might avoid false assertion simply to escape punishment, without this avoidance providing information about their doxastic states. Graham's concern, as Bruner summarizes it, is that deterrents fail to secure the connection between assertion and evidence because speakers frequently flout the norms of assertion, and because punishment is unstable and inconsistently applied. Bruner writes that deterrents "undergird a signaling system where speakers often flout the norms of assertion" and that "in many cases the evidential support is weak or nearly non-existent" (ibid.) and that, as Graham notes, "punishment is often fickle and comes with its own stability problems" (ibid., p. 5). Bruner proposes, instead, that internal sanctions can solve this evidential problem while functioning as receiver-independent costs (Bruner 2024, p. 8). As he explains, "what if sanctions are not at the behest of some other party, but instead originate from the speaker themselves. Internal sanctions of this kind are not too difficult to imagine. For instance, Alf might feel a pang of guilt when he asserts that the markets are favourable when he in fact knows they are in dire straits" (ibid.). When speakers have internalized norms of assertion, they experience psychological costs (guilt, shame) for dishonest assertion. These internal sanctions function as handicap-like mechanisms because the costs are receiver-independent rather than imposed by external punishment.

7. The breakdown of the metaphor: from deterministic costs to variable norms

Although Bruner's internal sanctions approach represents an interesting attempt to preserve the role of handicap-like mechanisms in human communication, it reveals a problem with applying signaling models to norms of assertion: the biological metaphor breaks down at the level of cost imposition certainty. In biological handicaps, signalling costs are imposed deterministically and automatically whenever the phenotype is expressed. A peacock with elaborate plumage always pays the metabolic costs (there is no probability involved, no circumstances under which the cost disappears). The cost imposition is a direct, invariable consequence of expressing the phenotype. This creates a "cost certainty problem" for handicap applications to assertion. Both external social sanctions and internal psychological sanctions impose costs

only probabilistically, not deterministically. External social sanctions depend on detection, community response, and enforcement (a false assertion might go unnoticed, unpunished, or face minimal consequences).

For instance: a mayoral candidate asserts during their campaign: "I will reduce property taxes by 15% in my first year if elected". They don't believe this is feasible given the city's budget constraints, but they know it will win votes. However, the detection mechanism fails because most voters don't closely monitor municipal budgets or have the expertise to verify the claim's plausibility during the campaign. Even when the candidate is elected and fails to deliver the promised tax reduction, the community response often proves inadequate. Many voters don't follow local politics closely enough to notice the broken promise, while others rationalize the failure by assuming economic conditions have changed. The enforcement mechanism also breaks down at multiple points. Even if some citizens recognize the broken promise, there's no formal mechanism to "punish" the politician beyond waiting for the next election.

Internal psychological sanctions do not solve the problem as they depend on the degree of norm internalization, situational factors, and psychological states (a speaker might feel no guilt or shame for false assertion in certain circumstances). For instance: a pharmaceutical executive testifies to the Parliament: "We had no knowledge of these side effects before drug release," knowing this is false – internal studies had documented the risks months earlier. The executive feels no guilt because years in corporate culture have normalized strategic communication as professional duty, making company loyalty supersede honesty norms. The high-stakes situation creates a context where dishonesty feels morally justified rather than shameful.

Although the handicap principle requires that costs be paid whenever the signal is produced (or at least, whenever a low-quality signaller attempts to produce a high-quality signal), human assertoric "costs" are probabilistic: they may or may not be imposed depending on circumstances, detection, internalization, and contextual factors. The probabilistic nature of human assertoric costs stems from how social norms are conceptualized within the social sciences and the humanities. The dominant conceptual framework (Bicchieri, 2006; Young, 1998, for instance) treats social norms as inherently variable, malleable, and context-dependent entities that resist the kind of stability required for deterministic cost mechanisms.

Here is why, in short, the handicap principle does not work in case one attempts to establish "social norms" as cost-imposing structures:

(1) Handicaps necessarily impose high cost for a set of signals ("assertions") to guarantee honesty.

(2) Social norms do not necessarily impose high cost on assertions, both as handicaps and deterrents.

(3) Therefore, social norms do not provide a handicap-like account of honesty as a supposed norm of assertion.

Clearly, the argument hangs on premise 2.

8. What are "social norms"?

Within the social sciences and humanities, social norms are categorically different from biological kinds. They are broadly constructed as culturally relative, and subject to interpretation and negotiation, and this assessment is also consistent with Austin and Searle. Since Green explicitly invoked Austin and Searle in his characterization of "social norms", I will first discuss in short what can be said about social norms based on their works.

Green's reconstruction of speech act theory reveals how both Austin and Searle see social norms in ways that support my claim. Green (2009, p. 148) describes "force conventionalism" – a view that "runs throughout Speech Acts and other works" – as the position that "for every speech act that is performed, there is some convention that will have been invoked in order to make that speech act occur." Austin explicitly acknowledged contextual dependency, noting that identical utterances have different normative effects: "Consider "I divorce you", said to a wife by her husband in a Christian country, and both being Christians rather than Mohammedans. In this case it might be said, nevertheless he has not (successfully) divorced her: we admit only some other verbal or non-verbal procedure" (Austin, 1962, p. 27). Searle's constitutive rules face the same problem through his "X counts as Y in context C" formula (Searle, 1969, pp. 34–35) – the crucial "context C" component reveals that institutional facts are always situation-relative, which undermines the consistent cost imposition required by biological handicap mechanisms.

Green himself agrees with the probabilistic character of cost-imposition when he concedes that "[e]xposure to risk does not, of course, guarantee harm: I could go all my life without a seatbelt or a scratch" (Green, 2009, p. 154). However, I will attempt to show that the lack of "guarantees" in such cases renders social norms incapable of functioning as stable, biology-style cost-imposition mechanisms at all and thereby prevents them from sustaining the analogy with the handicap principle. To further substantiate my claim, I will refer to several research papers that clearly demonstrate the general framework in which "social norms" are being discussed.

First, social norms are understood as historical phenomena that change over time, sometimes rapidly. Costs associated with false assertion in communities can shift dramatically as social attitudes evolve. What counts as a serious violation of norms of assertion in one period may be considered minor or acceptable in another. This temporal instability is documented in cross-cultural studies, with Harrington and Gelfand (2014) finding that cultural "tightness-looseness" varies dramatically across historical periods based on environmental threats and ecological pressures.

Second, cross-cultural studies reveal substantial differences in communication norms across societies. Some cultures value direct, explicit truth-telling while others prioritize politeness, face-saving, or indirect communication even when this conflicts with strict accuracy. Anthropological evidence shows that the social costs of false assertion vary dramatically across cultural contexts (what triggers severe sanctions in one culture may be considered normal conversational behaviour in another). Chen-Xia et al. (2023) found that Chinese participants showed significantly greater

emotional reactions to norm violations than British participants, despite rating the same behaviours as equally problematic. One could argue that cross-cultural variation diminishes when assertions concern high-stakes content. This is plausible. Consider false assertions about health risks or legal obligations, where cross-cultural convergence in sanctioning intuitively seems very likely. But it does not change the predicament because a mechanism that imposes costs conditionally, when stakes are high enough, in enough cultures, is still a probabilistic mechanism. The handicap principle requires costs that are invariant, not costs that arise only under favourable conditions.

Third, even within a single culture, the costs of false assertion vary dramatically across situations. Social sanctions for lying to one's spouse differ markedly from exaggerating one's qualifications in a job interview or making false claims in casual conversation, and professional environments often uphold stricter standards than informal ones. This variability aligns with Morris et al.'s view that "[…] norms are conceptualized as context-specific regulators of behavior rather than as traits […]" (Morris et al., 2015, p. 2). Moreover, they note that "norms can be identified with several objective features of a social environment" (ibid., p. 3), such as behavioural regularities, patterns of sanctioning, and institutional practices. Since these objective features differ across social settings, the norms they support, and therefore the social costs of violating them, also differ accordingly.

Fourth, different individuals within the same community may face different costs for identical assertoric violations. Factors like social status, relationship history, and personal reputation affect how communities respond to false assertions. High-status individuals may face lower social costs for false assertions, while marginalized individuals may face disproportionate sanctions for minor violations. Gavrilets et al. demonstrate that in social norm dynamics, "substantial inter-individual variability exists in all variables, reflecting individual psychological traits" (Gavrilets et al., 2024, p. 10) and that "attempts by an external authority to direct group behaviour can trigger an opposing behaviour (backfiring effect)" (ibid.).

Finally, social norms are not only variable but responsive to strategic behaviour. When speakers learn to game the system (avoiding detection, manipulating social responses, or compartmentalizing to avoid internal sanctions), the norms themselves often adapt in response. This creates a feedback loop where the very attempt to establish stable cost structures leads to their destabilization, not to mention the extent to which such activity is being facilitated by technologies for creation and distribution of assertoric content.

To summarize, social sciences have developed theoretical frameworks that emphasize the constructed, interpretive, and variable nature of social phenomena, including "social norms". This creates a problem when attempting to apply biological models that assume stable, law-like regularities, and renders stable, predictable cost-imposing structures impossible in principle, not just difficult in practice. Perhaps one could say that such an approach to social phenomena creates problems that go far beyond the philosophical adoption of the handicap principle, and one may not be wrong, but it is a risk I am willing to take.

9. Implications

The variability of social norms means that the differential cost structures required by handicap mechanisms as construed by Green cannot be maintained. Since costs are imposed probabilistically and vary across multiple dimensions, high-quality and low-quality signallers may face similar expected costs in many situations, eliminating the differential that makes handicap signalling work. As for Graham's deterrence model, the variability undermines the predictability and consistency required for effective deterrence. If potential violators cannot predict what sanctions they will face (or indeed, whether they will face any sanctions at all), deterrent effects are severely weakened. Bruner's focus on "psychological costs" such as guilt and shame still depends on internalized social norms, which are themselves subject to the same conceptual variability problems outlined above. Moreover, psychological states themselves represent yet another respect in which the biological metaphor fails. Unlike the automatic metabolic costs of maintaining peacock plumage, psychological states like guilt and shame are complex, interpretive phenomena that depend on cognitive appraisal, cultural meaning-making, and subjective experience. A peacock cannot choose to reinterpret the metabolic cost of fanning out its tail, but humans routinely reframe, rationalize, or cognitively restructure their psychological responses to norm violations. It might be objected that guilt and shame are themselves products of evolutionary selection, i.e., that individuals prone to these feelings were more adaptive because they avoided the long-term costs of dishonest communication. This may well be true at the ultimate level. It does not, however, affect the argument at the proximate level, which is where the handicap principle operates. A cost mechanism can have an evolutionary basis and still fail to impose costs reliably in the specific cases where assertoric communication takes place. The evolutionary origin of guilt does not guarantee that guilt is felt in any particular case.

To add to the mishaps, this cost certainty problem creates a contradiction analogous to what Rosen Lutskanov identifies as "self-undermining beliefs" ("самоснемащи се вярвания") in his analysis of Moore's paradox. According to Lutskanov, "[a] belief in a particular statement p [is] self-undermining if the presence of belief in p entails the presence of belief in the negation of p" ["Нека да наречем вярването в определено твърдение p самоснемащо се, ако наличието на вярване в p влече наличие на вярване в отрицанието на p."] (Lutskanov, 2014, p. 65). To illustrate the principle, Lutskanov describes an observer who sees someone confidently unlock and drive away a car, knowing that a skilled car thief recently escaped from prison. To the observer, the evidence supporting "this person owns the car" also supports "this person is the escaped thief," (ibid., p. 66) since a skilled thief would behave indistinguishably from an owner. Handicap-based accounts of assertion create an analogous predicament: for instance, observing confident public assertions (the evidence normally supporting "social norms impose reliable costs ensuring honesty") also supports the contrary belief that "social costs are unreliable enough to permit confident false assertion." When speakers recognize that social norms impose costs in a probabilistic manner, the confidence with which assertions are made becomes consistent with both honesty and strategic

deception, or, alternatively put, the same behavioral evidence that should demonstrate functioning cost-imposing mechanisms becomes evidence for their failure.

10. Conclusion

The biological handicap metaphor doesn't merely face some difficulties when applied to human assertion (as each metaphor does when its source and target domains are examined and compared in detail). In this case, the metaphor assumes interdisciplinary transfer of deterministic cost structures that are unavailable in social systems, as the latter necessarily produce probabilistic costs due to norm variability, precisely to explain how costs are imposed upon asserters. That is to say, the transfer of cost-imposing mechanisms from biology to philosophy is the very reason the metaphor was created and since the transfer is impossible and not merely in degree impossible, the metaphor is meaningless. We are thus not discovering that human assertions work through mechanisms like biological handicaps; rather, we are using handicap imagery to think about human communication while ignoring the fundamental differences that make the comparison inaccurate.

This pattern of normative failure due to descriptive inadequacy reflects a broader methodological challenge in contemporary philosophy. A parallel diagnostic emerges in Madelaine Angelova-Elchinova's critique of traditional belief theories, where she demonstrates that normative accounts fail when they rest on empirically inadequate descriptive foundations. As Angelova-Elchinova shows, appealing to neuroscientific evidence about belief formation, traditional normative theories assume that "beliefs are reflective, b) conscious doxastic attitudes towards a proposition," but this "Transcendental View about Belief" is "normatively inadequate, because it is descriptively false" (Elchinova, 2024, p. 117). While Angelova-Elchinova examines belief formation at the cognitive level and I discuss the failed application of social norms as cost-imposing structures, both cases illustrate the same methodological principle: when philosophical theories assume structural features that empirical research shows to be absent or fundamentally different, the resulting normative frameworks become systematically inadequate.

Postscript: addressing two predictable objections

A natural worry about the argument developed so far is that it relies on an artificially stark contrast between biological determinism and social variability. Two objections are particularly salient.

1."Social norms can become rigid and effectively deterministic."

One might object that the variability of social norms has been overstated. Norms embedded in long-standing institutions or "tight" cultures may appear to impose sanctions with sufficient regularity to mimic the cost patterns of handicap mechanisms, legislation being the clearest type of such mimicry. Although the legal integration of

social norms is indeed a salient feature of the human environment, there are significant caveats that do support my claim. To put it succinctly, the empirical literature on norms emphasizes that their operation is expectation-driven and historically fragile rather than cost-deterministic.

For instance, Cristina Bicchieri insists that norm compliance depends on conditional expectations about others' behaviour and attitudes. Discussing fairness norms, she writes that "conforming to a social norm of fairness is different from endorsing fairness: I may conform to a fairness rule if I have beliefs about what others who are relevant to me do and approve of, but I may care little about fairness per se" (Bicchieri, 2017, pp. 67). The preference to conform is conditional on empirical and normative expectations; when those expectations shift, the conditional preference can collapse. The same structure reappears in her analysis of pluralistic ignorance. She notes that agents can come to "embrace a norm that is perceived as widely supported" purely in order to reduce the discrepancy between their private and public attitudes (Bicchieri, 2006, p. 194). Over time such internalization may turn a misperceived norm into a genuinely shared one, but only through a dynamic process of changing expectations and self-understanding (ibid.). This is a mechanism of norm entrenchment, but it is not a mechanism of deterministic cost imposition: the same pattern of behaviour can be supported by different underlying expectations, and small informational changes can tip the system into a different equilibrium. In her analysis of tax compliance, Margaret Levi notes that the "breakdown of quasi-voluntary" compliance is signalled not merely by rising rates of formal objection but by broader shifts in taxpayers' beliefs about fairness and state performance (Levi, 1988, pp. 168-169). Whether sanctions are accepted as legitimate and whether they reliably motivate behaviour depends on context, perceived justice, and expectations, not simply on the fact that a formal penalty exists. Finally, Brennan et al. argue that the desire to avoid formal sanctions "is, in fact, rather limited" as an explanation of why we obey the law: "We obey the law instead because of our 'normative commitments'" (Brennan et al., 2013, pp. 229–230). If obedience to legal rules is typically grounded in moralized commitment rather than straightforward sanction-avoidance, the idea that legal or social rules can function as deterministic cost-imposing structures becomes even less plausible.

These accounts suggest that even where norms are stable, their stability is mediated by expectations, internalization, legitimacy, and perceived fairness. None of these mechanisms guarantee that violations will be met with fixed, unavoidable costs. This type of stability is statistical and historical, not mechanistic in the way required by handicap models.

2."Biological signalling is not fully deterministic either, so the contrast is overstated."

A second objection grants that social norms are variable but denies that biological signalling is as deterministic as I have suggested. After all, biological systems are themselves subject to environmental variation. The contrast, the objection continues, may therefore be exaggerated. Indeed, an absolutist claim about the domain of biology being deterministic is hard to defend and I am not going to die on that hill. The

relevant question, however, is not whether biological systems are perfectly deterministic in every respect, as I did not claim anything along these lines, but whether the cost structures posited by handicap models are qualitatively or degree-wise more constrained than social costs. All that my claim requires is that they are constrained to a degree which cannot be used to characterize social costs and, on this point, biological literature is clear enough.

Grafen's model (Grafen, 1990, pp. 517–533), defines the signalling system exclusively in terms of biological variables – true quality q, advertising level a, perceived quality p, and the fitness functions $w(a, p, q)$ and $D(q, p)$. Equilibrium honesty is derived from fitness maximization under the condition that the marginal cost of signalling varies monotonically with underlying quality. No social, psychological, or normative parameters play any role in the analysis, nor does the model appeal to belief states, shared expectations, reputational mechanisms, or sanctioning practices. All mechanisms necessary for honest signalling are intrinsic properties of organisms' cost-quality structure as specified by the model. Maynard Smith and Harper's work emphasizes that the handicap principle is only one among several mechanisms that can stabilize honest signalling. They distinguish "strategic costs" that are "necessary in order to ensure reliability and prevent 'lying'" from mere "efficacy costs" of message transmission and note that models by Enquist, Pomiankowski, and Grafen "have shown that there are indeed contexts in which the reliability of signals can be ensured by costs" (Maynard Smith and Harper, 2003, p. 30). They summarize the core condition as follows: "The requirement for evolutionary stability of costly signalling is that the fitness cost of a given signal must be relatively lower for signallers of high 'quality'" (Maynard Smith and Harper, 2003, p. 31). They also emphasize that there are alternative mechanisms for reliable signalling: "there seem to be three alternative explanations for the reliability of signalling systems – the handicap principle, common interest, and indices of quality" (Maynard Smith and Harper, 2003, p. 9). In all three cases, reliability is secured by features of the biological system (differential costs, overlapping interests, or hard constraints) rather than by socially mediated sanctions.

These are but a few of many examples demonstrating that biologists retain the idea that whatever cost or constraint sustains honesty is grounded in physiology, ecology, or hard informational limits. The relevant "cost" is not something that can be evaded by reinterpretation, renegotiated by changing expectations, or by gaming the system. It is built into what organisms can do, not into what they or their peers think they can do and/or should do. By contrast, the social mechanisms invoked in norm-based and hybrid accounts of assertion are inherently revisable. Expectations can be revised, internal sanctions can be weakened or overridden, and formal sanctions can be ignored, nullified, or deemed illegitimate. These are not defects of the empirical record but structural features of how social normativity in modern societies is conceptualized.

References

Angelova-Elchinova, M. (2024). Norms of belief and non-propositional primal beliefs. *Balkan Journal of Philosophy, 16*(2), 117–130.

Austin, J. L. (1962). *How to do things with words*. Oxford University Press.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). *Explaining norms*. Oxford University Press.

Bruner, J. P. (2024). Assertions: Deterrent or handicap? A reply to Graham (2020). *Episteme, 21*(1), 1–11.

Chen-Xia, L., Wang, Y., & Zhang, M. (2023). Cultural variations in perceptions and reactions to social norm transgressions. *Frontiers in Psychology, 14*, Article 1243955.

Gavrilets, S., Tverskoi, D., & Sánchez, A. (2024). The interplay of social identity and norm psychology in the evolution of human groups. *Philosophical Transactions of the Royal Society B: Biological Sciences, 379*.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology, 144*(4), 517–546.

Graham, P. J. (2020). Assertions, handicaps, and social norms. *Episteme, 17*(3), 349–363.

Green, M. S. (2009). Speech acts, the handicap principle and the expression of psychological states. *Mind & Language, 24*(2), 139–163.

Harrington, J. R., & Gelfand, M. J. (2014). Tightness–looseness across the 50 United States. *Proceedings of the National Academy of Sciences, 111*(22), 7990–7995.

Levi, M. (1988). *Of rule and revenue*. University of California Press.

Lutskanov, R. (2014). Paradoksat na Mur i logikata na obosnovavaneto. *Filosofski Alternativi, 6*, 60–68.

Maynard Smith, J., & Harper, D. (2003). *Animal signals*. Oxford University Press.

Morris, M. W., Hong, Y. Y., Chiu, C. Y., & Liu, Z. (2015). Normology: Integrating insights about social norms to understand cultural dynamics. *Organizational Behavior and Human Decision Processes, 129*, 1–13.

Penn, D. J., & Számadó, S. (2020). The handicap principle: How an erroneous hypothesis became a scientific principle. *Biological Reviews, 95*(1), 267–290.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

Young, H. P. (1998). *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press.

Zahavi, A. (1975). Mate selection: A selection for a handicap. *Journal of Theoretical Biology, 53*(1), 205–214.

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 211 – 222

EMBODIMENT, PLAYFULNESS, AND HUMANIZATION

David-Augustin Mândruț*

Abstract: This paper aims to outline the relation between embodiment, playfulness, and humanization in the case of mother-infant interactions. To accomplish this task, I will firstly draw on the works of Marc Richir, since he has provided a full-fledged phenomenological account of key concepts pertaining to the psychoanalytical theory of British psychoanalyst Donald Winnicott. Even though this essay will be situated mostly in the framework of Richir's thought, I will also draw on different phenomenological authors, such as Eugen Fink or Paul Ricoeur, whose ideas can strengthen my initial claim. The process of humanization, as described by Richir entails the entire range of playful gestures, which reveal our embodied being. Therefore, interactions such as the exchange of gazes, mother-infant vocalization and touch will be analyzed. Finally, this article attempts to introduce a novel reading on the notion of resistance, to reveal its ethical implications regarding the mother-infant scenario. The process of humanization does not involve only how the infant commences to learn about aspects concerning what it means to be human, but also it entails the ethical relationship to others, henceforth, the notion of empathy will be also addressed.

Keywords: embodiment, play, affection, sublime, reality.

1. Introduction

This essay engages with the thought of Marc Richir, to disclose the basic tenets of the relation between embodiment, play, and the process of humanization. First, I will be arguing that embodiment is the main feature of the infant's play, which permits the discovery of the shared world in its fullness (his/her *Leibkörper*, the first empathy toward the other, external reality etc.). Following the observations of Marc Richir, who himself has drawn on Donald Winnicott's work, the infant's playfulness renders possible the discovery of the body, understood as *Leibkörper*, namely both as lived and experienced and as an object. Moreover, by virtue of the body, play becomes possible. Eugen Fink has stressed in several ways the pivotal role of embodiment in any sort of play-activity, while Marc Richir discussed the playfulness of interactions such as face-to-face exchanges (*l'échange des regards*), but he also alluded to the playfulness of babbling in the case of these face-to-face scenarios. Following the notion of play, whose structure resembles the illusion-model,[1] proposed by Joona Taipale's reading of Winnicott, it could be argued that the separation of the interior and exterior (worlds),

[1]Joona Taipale compares and then distinguishes between Winnicott's model of play and Fink's one, thus he terms the former the illusion-model, while the second could be designated as the hybrid model (Taipale, 2021). However, a discussion concerning the key differences between these two accounts will not represent an aim of this paper.

*Babeș-Bolyai University of Cluj-Napoca, Faculty of History and Philosophy, Center for Applied Philosophy; ORCID: 0009-0009-8644-3243; davidmandrut@gmail.com

©2026 David-Augustin Mândruț. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

is rendered possible by the resistance of the infant's body, implicated in his/her play-activity. Thus, the body will be considered under two basic aspects of its existence, i.e. as resistance and as the possibility of self-affection[2]; however, the issue of hetero-affection will not be overlooked by our present study[3]. The main claim of this paper is that the close interrelation between embodiment and play is the *via regia* towards the process of humanization. Therefore, the present article analyzes two possible approaches to the problem of playing, namely the one implying the infant playing alone, self-affecting himself/herself, and the situation in which the other is implicated in this sort of discovery, by virtue of the interactions mentioned above. Finally, by tracing the relation between embodied playing and the process of humanization, this study aims to prove that Winnicott's wording about playing and reality can be read as a sort of passage from embodied playing to several aspects of shared reality, which also involve our emphatic relation to others and to the disclosure of different aspects of the world.

The relation between playing and embodiment has been analyzed several times by different authors, both in recent literature and in canonical texts pertaining to the history of phenomenology, while the relation between embodied playing and the process of humanization has been left aside or overlooked. This paper aims precisely at an investigation of the metamorphoses of embodied playing throughout the infant's process of humanization. Playing can be broadly understood as any activity in which infants reach creatively towards several aspects of reality, including the relation to others or to themselves, thus rendering possible various aspects of the world, even the discovery of one's body. I will therefore gather the basic tenets of Marc Richir's theory of the process of humanization (Richir, 2010, p. 31), at the same time corroborating it with different accounts of the relation between play and embodiment.

Throughout his later work, starting with *Phantasia, Imagination, Affectivité* (2004), Richir inquired into the works of Donald Winnicott, in an original attempt to describe the stages through which the infant passes to accomplish the transition from, so to say, drive to presence[4]. Shortly put, this transition is rendered possible by several phenomena pertaining to the infant's relation to others, such as the "moment" of sublime, and the symbolic institution of language. While Winnicott considered embodiment in the context of drives (Winnicott, 2009, pp. 69–70), Eugen Fink regarded embodiment as the key element of playing, a *leitmotif* (Fink, 2016, pp. 227–228). Therefore, this paper will be divided into several sections, each investigating a key aspect of the relation between mother and infant, pertaining to the most archaic register of phenomenality, in which embodied play is involved. Thus, I will start with the archaic register of phenomenality in which physical contact is involved, then I will turn to the exchange of regards, to finish with the so-called exchange of babbling, or what could be termed as mutual vocalization.

[2]We are thereby introducing the two facets of the *Leibkörper*, the subjective, lived one, and the objective one.

[3]For a crucial discussion concerning the difference between self-affection, auto-affection, and hetero-affection, see Barbaras (2021, pp. 284–285)

[4]We are alluding to the works of Henri Maldiney, wherein he analyzed the complicated, yet essential interrelatedness between drive and presence (Maldiney, 1991).

It can already be noticed that at each key point of the archaic lifeworld of the infant, the body is already operative. This functioning body is essentially playful for the infant, although one cannot neglect the key roles played by the *chora* in Richir's terms or the *primordial Leib*, our paper will consider the body under the heading of what Winnicott has called the limiting membrane, i.e. the body construed as displaying an inside and an outside, namely as individuated (Winnicott, 2009, p. 3). Somehow, even though the mother or the infant engages first in this sort of interaction, play is a pervasive phenomenon, which pertains to our bodily being. Before concluding our essay, a subchapter dedicated to the issue of the experience of the external and internal world, in which resistance and effort are interrogated, will be addressed. Finally, this paper considers the process of humanization as being dependent upon the notion of embodied playfulness.

Even though these interactions pertain to the so-called "immemorial", i.e. the most archaic stage of the human being, they nonetheless set up the infant's well-being, as it will be shown. The process of humanization designates how the infant acquires or discloses certain aspects of what it means to be human, thus, the becoming-more-human is rendered possible by his interactions with the mother, in the case of which symbolic institution is paramount. Henceforth, embodied playfulness becomes the *sine qua non* of the process of humanization, whereby these early face-to-face interactions which involve contact (even at a distance)[5], render possible the infant's empathy toward the other. The first empathy becomes thus the first mark of the accomplishment of the passage from drive to presence. The passage from drive to presence does not designate a sort of, so to say, natural transition, but rather it is dependent from the very beginning on the relation with the mother, the concepts of the "moment" of sublime and of symbolical institution being pivotal for any inquiry into the process of humanization whatsoever. As we are going to notice, there is a tension between self-affection and hetero-affection, whereas both are somehow involved in the face-to-face interactions of mother and infant. Nonetheless, both the affecting of oneself and of being affected by the other represent peculiar ways in which the comprehension of our embodiment unfolds.

2. The playfulness of touch

The first layer of our analysis tackles the question concerning the reciprocity of touching and being touched. From the very beginning, we have to employ the notion of self-affection and hetero-affection, to understand the reversibility which is at play in this sort of interaction. I insist on demarcating the notion of self-affection from Michel Henry's auto-affection (Henry, 1973, p. 234), since the latter's notion was conceived in a wider framework of thought pertaining to the phenomenology of life[6]. While for Michel Henry, auto-affection designates the radical immanence and interiority of our affective life, independent of the ekstasis of the world, the notion of self-affection to be employed by us refers simply to the infant's creative reaching out to his/her body due to different modalities of our corporal being, such as hearing one's cry, etc. For

[5]In the case of the exchange of regards or babbling, for example.

[6]For a disambiguation of this distinction, see for example Sheets-Johnstone (2006).

the part being, I will use the notion of self-affection to designate the sort of behavior the infant exhibits towards his/her body, such as playfully touching one hand with another, etc. But as we will quickly notice, self-affection does not necessarily imply only the chiasm of the touching body and the touched body (Merleau-Ponty, 1968, p. 130), but it can go across several modalities, such as Winnicott's famous example of the infant's cry, which was prolonged by him/her, to feel the pleasure of a musical note (Winnicott, 2009, p. 92). Thus, using the vocal capacity to cry or to shout, the infant self-affected himself/herself in another modality, such as hearing, etc.

From the very commencement of their lives, infants are held by their mother, and this sort of holding does not only provide physical comfort and warmth, but also, it designates an opportunity for a sort of exchange of touches, where the mother caresses the infant's skin, and the infant might even try to touch the mother, as in Winnicott's famous example of mother-infant reciprocity (Winnicott, 1989, p. 251). A very useful remark pertaining to the framework of Sartre's ontological phenomenology, in which caressing the other's skin reveals his/her embodied freedom (Sartre, 2003, pp. 411–412), thus constituting their bodily being as displaying and exterior and an interior (Winnicott's limiting membrane). Therefore, the reciprocity of touching and being touched represents the passage from *chora* and the *primordial Leib* to the body as being spatially individuated, with an exterior and an interior of its own. By the same token, Renaud Barbaras employs also the case of caress, to show how the reciprocity and mutual reverberation between touching and moving enhance one another in a sort of process of individuation of different modalities of experience (Barbaras, 2011, p. 69).

With respect to the situation of the mother-infant scenario of mutuality, Winnicott reports that the infant tries to put his/her finger into the mother's mouth to reciprocate. This behavior already indicates a sort of *"perceptual" phantasia* as Richir would insist, but we will limit ourselves to the investigation of this corporeal relation, leaving aside the possibility of presenting the pivotal role played by *"perceptual" phantasia*; thus, we will rather be interested in the possibility of exchanging touches. As Francine Wynn has demonstrated, not only does the mother hold her baby, but also the baby holds his/her mother. Therefore, Wynn advocated for a sort of embodied and chiasmic relation between mother and infant (Wynn, 1997, pp. 258–259). Not only do the mother's touches and caresses build up the infant's feeling of security and trust in the world, as Fazakas and Gozé have proved in a phenomenological and transcendental manner (Fazakas & Gozé, 2020, p. 175), but also, this sort of reciprocity enhances and deepens the mother's skills and abilities towards her baby. As we have already stated, employing the vocabulary of Sartre, touch, whose most important manifestation is the caress of one's skin, demarcates the other's freedom by instituting the difference between exterior and interior. Therefore, by virtue of the validation coming from the part of the infant, i.e. that everything worked well in the interaction, the mother's skills are improved. Moreover, it could be insisted that the infant and the mother find themselves in a process of trial and error, in which they have to attune and reattune permanently (Stern, 1998, p. 138), at the level of their bodily beings.

We can also stress upon the rather and seemingly asymmetrical aspect of the mother-infant touch interaction, whereas usually the mother initiates this interaction, but nonetheless, the infant might echo it, creating a sort of bodily resonance, which contributes to the process of bonding between them (Patočka, 2016, p. 166). Therefore, not only does the infant initiate the interaction with the mother by virtue of his/her playful spontaneity, but the mother's response strengthens the relation between the two of them, consolidating the bonding scenario. While the asymmetrical, yet usually reciprocal relation between mother and infant concerning touch might seem at first something evident, Marc Richir provided several useful examples to prescribe the first spatial relations which the infant establishes between his different body parts. For example, Richir emphasized the gesture of sucking one's thumb, as a sort of "opportunity" for differentiating the interior from the exterior (Richir, 2006, pp. 275–276), and thus, establishing a correlation between inner and outer worlds. Another very useful example of Richir's indicates towards the chiasm between the infant's *Leib* and *Körper*, which results in the experience of *Leibkörper*, this being rendered possible by the infant's playful self-affection of his/her own body by touch, for example the hand touching the other hand etc. (Richir, 2015, p. 244). As we have stated above, we will not concern ourselves for the time being with the most archaic register of spatiality, yet the mentioning of the *primordial Leib* and the *chora* might strengthen our argumentation. Richir borrows the notion of *chora* from Plato, which represents the receptacle holding the infant. *Chora* can even resemble the maternal lap, a genuine prototype of the Earth that does not move (Richir, 2006, p. 268).

Returning to the to-and-fro play between self-affection and hetero-affection, i.e. affecting oneself or being affected by others, it must be emphasized that even the earliest gestures of the infant towards the other and towards his/her own self have this peculiar function of establishing what pertains to the internal realm and what belongs to the external world. As Fazakas and Gozé have stressed, it is first by touch that the infant discovers the Earth, in a phenomenological sense (Fazakas & Gozé, 2020, p. 182). Therefore, thanks to the "mutual reverberation" between touch and movement, as we have glossed this phenomenon following Barbaras, the Earth is explored and tested perpetually. Renaud Barbaras considered Merleau-Ponty's notion of flesh to resemble the Husserlian Earth (Barbaras, 2021, p. 68), thus the correlation drawn by Patočka between the relation to the Thou and being an "earthling", i.e. the fact of being bound to the Earth, proves to be crucial for the phenomenological analysis of the infant's world (Patočka, 1998, p. 149).

3. The exchange of gazes

The other modality to be presented is the visual one, which finds its most important moment in the exchange of gazes between mother and infant. For the part being, I will not inquire into the exchange of *"perceptual" phantasia*, involved in this sort of crossing of regards, but rather, I will analyze its playful character and rhythmical modulation, which are rendered possible by the "moment" of sublime. The "moment" of sublime is the very instant in which humanity is engaged, as Richir holds, but nonetheless, despite being the very first human or even interhuman experience, the "moment"

of sublime represents the tension between the two poles of this feeling, described by Kant, these being fear and pleasure. Nevertheless, this moment can be coupled with symbolic institution to reveal how the process of humanization is advanced further during the essential stages through which the infant passes, due to what Richir has called the symbolic institution of language. The "moment" of the sublime can resemble maternal love, as Prášek has shown (Prášek, 2021, p. 67), yet this statement must be understood as an excess of affectivity whereby the mother's role is pivotal, since she cannot let the infant fall into the abyss (Fazakas & Gozé, 2020, p. 184), and thus, her presence and the atmosphere she provides for the infant are essential for the child's capacity for distinguishing a fortunate encounter with this excess of affectivity, and an unfortunate one. Although there exist certain similarities between Kant's theory of the sublime and Richir's phenomenological sublime, the latter pertains to another type of affectivity, which does not correspond either to the mathematical or dynamic sublime, but rather involves this excess of affectivity, i.e. the overflowing maternal love, whose main virtue is that of opening up the space of the encounter with what Richir has glossed as the absolute transcendence (Richir, 2010). Once the baby has felt the mother's infinite love, his/her infinite aspiration of affectivity is awakened, too. This infinite aspiration of affectivity should be contrasted with the nostalgia of the immemorial, which Richir discusses in his last books (Richir, 2016, pp. 181–182), We once again encounter the implicit notion of tension, generated by the "moment" of sublime, between the infinite aspiration of affectivity, and the immemorial "call" of nostalgia. Therefore, if we construe maternal love as a species of the "moment" of sublime, the tension between future (immature) and the past (immemorial) might shed light on the very relation between the infinite aspiration and nostalgia, these being rendered on the affective level of our being. Therefore, following Richir's account, one can notice this perpetual tension between the infinite aspiration of affectivity, which moves forward, and nostalgia, which aims towards the transcendental past (Richir, 1987), one attempting to stretch toward the immature future, and the other toward the immemorial past. But what exactly has the "moment" of sublime to offer to the exchange of gazes? The infant and mother turn to each other and cross sights, which permits the infant's gaze to be fixed, and thus to institute two "absolute here" – his/her and the mother's. This represents the very commencement of, so to say, interpersonal spatiality. We can think of the many examples provided by Daniel Stern concerning the peek-a-boo game, in which object constancy was in the process of being built. The exchange of gazes can even be considered as the prototype of the peek-a-boo playful interaction, as we can suggest. Richir did not concern himself with how the infant builds up object permanence or constancy, yet he provided a thorough analysis of Winnicott's text on the use of an object, in which this theme is addressed. Although this commentary, to be found at the end of *Phantasia, Imagination, Affectivité* is very insightful (2004, p. 497), its analysis will not represent an activity of this paper. In this specific scenario, in which the mother lovingly glances at the infant, smiling for example (Patočka, 2022, p. 129), the infant is hetero-affected, receiving the positive "moment" of sublime. As Winnicott has proved, if this hetero-affection functions at the beginning like a mirror, in which the infant sees himself, the mirror thus provides something back to the infant, something which has belonged to him/her all along.

Thereafter, the infant will not be confused when regarding a mirror (Winnicott, 2009, p. 149). Once again, this perpetual exchange of gazes ends up in a sort of bodily resonance, due to its echoing function, which once again builds up the infant's trust in the mother, which leads to continuous bonding. Thanks to the echoing function and to resonance, distance is somehow instituted between mother and infant, which again points to the very emergence of human spatiality[7]. Moreover, Husserl himself has already emphasized how the process of mirroring between mother and infant represents the way in which their interaction unfolds, this might even trigger the very first manifestation of language (Husserl, 1973, p. 608).

For now, we have insisted upon the playfulness of the exchange of gazes, even considering it to be one of the early forms of interaction between mother and infant. We ought to add that this particular type of interaction is also a rhythmical one, because the echoing and resonance are only achieved by virtue of the infant and mother's synchronization of their durations, or of their tensions of duration[8] (Schutz, 1976, p. 159). These tensions of duration could be designated as the rhythms of existence, since tensions involve a certain intensity. The tension of duration can be equated with rhythm, understood not necessarily as a pace-movement, but rather as the specific tension into which the subject finds itself. Thus, the exchange of gazes involves a sort of mutual focus, so the mother and infant have to share the same tensions of duration, which results in their synchronization. Moreover, Schutz's theory of musicality can be compared with Stern's affect attunement, since both theories entail that the fluxes of duration of both mother and infant have to be synchronized for the interaction to be a satisfactory one.

Henceforth, affectivity, provided by the sublime, is the second component which renders possible the exchange of gazes, and correlatively, the process of humanization. Richir has insisted upon the sublime feature involved in this exchange and perhaps we can now see clearly why. The sublime appears to us as carrying along a sort of surprise, which seizes our entire being, even the body. Therefore, even though the interaction just described is called the exchange of gazes, the feature of embodiment is involved all along. One partial conclusion to be drawn already is that we are from the very beginning embodied beings, and that our way of handling existence at the very beginning involves playfulness, i.e. the creative reaching out. Not only does embodiment require playfulness at the early stages of our existence, or playfulness requires the body to become operative, but also, their interdependence is so essential that we can just consider, as Shaun Gallager does, that these types of interaction are somehow present throughout our entire lives (Gallagher, 2020, p. 2020). Without appealing to the Hegelian dialectic of suppression and conservation, Gallagher showed how both primary and secondary intersubjectivity are built in our daily face-to-face

[7]Richir's reading of Winnicott sympathizes with the latter's statement that the infant is initially in a state of fusion with the mother, and only afterwards do distance and differentiation occur.

[8]Schutz has argued for the interrelatedness between the tensions of duration and the attention to life, following Bergson's thematization. We can consider as well that in the case of the mother-infant scenario, the synchronization between the two of them involves a sort of joint attention (Gallagher, 2020, p. 106).

dyadic interactions with others. We already provided a hint towards the relation between embodiment and playfulness, but we have yet another interaction to analyze. We are now ready to move towards the exchange of babbling, because as we can notice, the mother not only glances at her infant in a loving manner, but also, she talks to him/her, at the very beginning, in a sort of babbling.

4. The first musical notes

This subchapter of our analysis consists in the investigation of the exchange of babbling between mother and infant. For the infant, babbling is a very pleasurable activity, and the response coming from the mother's part encourages the infant to babble more, and to finally reach a word. We can identify babbling to lie at the cornerstone of language, this being rendered possible by the symbolic institution of language. Nonetheless, Richir is very attentive from the very outset in defining babbling as a sort of proto-musicality, such being the case of language, too (Richir, 2010, p. 21). Thus, corroborating the theories of Schutz, Richir and Stern, it can be argued that the forerunner of language is human musicality. Richir envisages this sort of interaction between the babbling of the infant and that of the mother as being coupled with the exchange of gazes. Henceforth, to use Wynn's phrase, between mother and infant there emerges a genuine symphony of gestures (Wynn, 1997, p. 259). Although this type of interaction might seem rather simple, its phenomenological characteristics ought to be outlined further. As we stated before, drawing on the example provided by Winnicott, the infant finds pleasure in extending the duration of his/her cry, because finally, he/she causes a musical note to emerge. We could consider the babbling of the infant to be yet another example of playful self-affection, since his/her babbling is already rhythmical, therefore musical, as Richir stresses. We can already point toward the idea that the infant feels addressed by the mother during the exchange of babbling, because the mother looks into his/her eyes, and thus, fixes his/her gaze. We have employed the term self-affection to describe the infant's babbling, but also the moment when he/she manages to touch his/her body, for example the hand on the arm, to insist once again on what Richir has called the contact between self and self (Richir, 2010, p. 31). The contact between self and self occurs prior to what Michel Henry has called auto-affection, because, as Richir argues, a gap (*écart*) is necessary between the two components of the self for auto-affection to occur (Prášek, 2021, p. 68). Therefore, Richir even considers the self's structure to be twofold, their contact being rendered possible by the "moment" of sublime, i.e. by hetero-affection.

Moreover, Richir develops the idea that the infant who babbles has already a sort of spatial landmark, because the intertwining between the exchange of gazes and that of babbling, already presupposes the absolute here, or the zero point of orientation as Husserl suggests. The very interesting, yet evident difference between Marc Richir's approach and, for example, that of child psychologist Daniel Stern, consists in the way in which they thematize early mother-infant interactions. While Marc Richir advocates for the perspective which considers the mother-infant relation as a sort of reciprocal rapport, in which each modality of the infant is responded to by that of the mother, Daniel Stern insists on the insufficiency of such reciprocal comportments.

For Daniel Stern, affect attunement has to be a cross-modal interaction, in which the infant makes a hand gesture for example, while the mother encourages vocally that gesture (Stern, 1998, p. 138). Furthermore, while for both authors the body was pivotal for these types of interactions, Richir stressed more the issue of *"perceptual" phantasia*, while for Daniel Stern, imagination seems to have been overlooked, since he discusses the developmental stages through which the infant passes, which must not be mistaken with what Richir has called the transcendental genesis of subjectivity. Once again, Stern was interested in the stages through which the infant passes and correlatively to the maturation processes, while Richir was rather concerned with the phenomenological life of the infant, entailing the phenomena described above, such as the exchange of gazes and *"perpetual" phantasia* or the "moment" of sublime. As Richir points out, the phenomena taking place in the case of transcendental genesis can be simultaneous, while in the case of child development, the specific phenomena of maturation are successive (Richir, 2010, p. 197).

5. Resistance and effort: Between inner and outer worlds

Now that we have emphasized these sorts of interactions, we are ready to move toward the part of our paper which tackles adjacently the contributions of Eugen Fink and Paul Ricoeur. Marc Richir has himself drawn on Fink concerning several aspects of transcendental phenomenology, but not on his theory of play. Even though both authors have borrowed the notion of *phantasia* from Husserl, their manner of applying it to philosophical subjects differs radically. Eugen Fink considered that imagination fuses with reality, and this gives rise to the *playworld*, while Richir somehow overlooked the problem of play, and rather focused on the notion of *phantasia* in the case of interpersonal relations, such as the exchange of gazes. In the case of this analysis, bearing the three moments of the Richirian analysis in mind, we will move toward an account of the *playworld*, in which resistance is involved.

Therefore, resistance should be defined in a twofold manner. First, it involves the resistance of so-called external reality, and second, it entails the resistance towards one's own body, for example, when fatigue emerges. The first type of resistance will be conceptualized according to Winnicott's thematization, while the second type will follow Ricoeur's description from *The Voluntary and the Involuntary* (Ricoeur, 1966, pp. 310–311). While Winnicott described the resistance of the object as the first attempt to lance an act of empathy towards the other, Ricoeur stressed how the resistance of one's own body gives rise to the feeling of effort. Following Taipale's proposal of the illusion-model theory of Winnicott's framework, already mentioned (Taipale, 2021, p. 202), in which throughout the play activity, imagination is merged with reality from the very beginning, it could be argued that both the resistance of the external world, and that of the body play the peculiar role of demarcating the external from the internal. For example, during the infant's play, external reality might oppose resistance vis-à-vis the infant's movements, thus causing the body to feel the effort of attempting to reach something. Therefore, and following Ricoeur's argument, effort emerges when the infant's will or desire is not satisfied at the exact moment of need, thus the infant somehow leaves the illusion-model of the *playworld* temporarily. In a

rather phenomenological fashion, during the infant's playing, the "I can" of Husserl's becomes omnipotent due to the imaginative contents of his/her mind, whereas when encountering the resistance of external reality, and thus causing effort to emerge, the "I can" loses its illusory omnipotence and finds itself somehow frustrated due to the incapacity to satisfy the imaginary desire, which fueled the play activity (Husserl, 1989, p. 277). As Heidegger, for example, has noticed in *Being and Time*, Dasein finds out that it is not the master of the world, when it encounters the resistance of the world as such (Heidegger, 1962, p. 407).

Likewise, the first act of empathy towards the mother can be seen as involving physical contact. Winnicott describes the peculiar situation in which the infant bites the mother's breast during feeding. The mother will lance probably an interjection such as "Auch!", and this will awaken the first empathy of the infant toward the mother, because the infant will acknowledge that the mother is vulnerable, like every human being (Winnicott, 2009, p. 115). Moreover, not only will he/she recognize that the mother is as vulnerable as every other human being, but also the infant will accomplish the passage from the mother seen as an object to the mother seen as an environment, by virtue of her capacity to resist the attack of the infant's bite (Winnicott, 1965, pp. 75–76). Therefore, by virtue of the mother's survival of the infant's attack, she will become independent of the infant's mind, and somehow, she will be regarded as possessing a life of her own, namely independence. Even though Winnicott has described this process in a somewhat impersonal manner, and Richir has drawn on him, without insisting on the mother's vulnerability and the emergence of empathy, both had something very important to say about the way in which the interior world is demarcated from the exterior one. Our example provided above might even represent a phenomenological reading of the infant's emergent empathy and to his/her first relation towards external reality, thus establishing object constancy. Moreover, we could also argue that the infant's "attack" develops in two different, yet related moments, i.e. that of effort, namely the infant's attempt to bite the mother, and the mother's survival, i.e. resistance. This sort of interaction has the essential function of demarcating the interior from the exterior and thus permitting the infant to launch an act of empathy, by which external shared reality is achieved. This could be the very way in which the infant comes out of the state of fusion with the mother, namely by setting her at a distance, and regarding her as something independent of his/her mind. Once again, this sort of relationship helps to strengthen the spatial relation between mother and infant by fortifying two "absolute here", that of the infant and that of the mother, the latter being now a separate, not-me being for the child.

6. Conclusion

In this paper we have attempted to lay out the basic tenets of the process of humanization in a threefold manner. The three modalities that have been analyzed were meant to demonstrate how the playfulness of embodiment operates already from the very beginnings of the mother-infant relation. Therefore, we can conclude that the infant's life commences with the early contacts with the mother, which are tactile ones. Then, it follows the exchange of gazes, and its complementary interaction,

which is precisely the exchange of babbles. Finally, we have provided some hints and clues towards the first spatial demarcations of the infant's life, which consisted in the way in which he/she can differentiate between inner and outer worlds, by virtue of the experience of resistance and effort. It could be argued once more, alongside Winnicott, that play is the *via regia* towards reality. This statement does not mean that the infant is first in an isolated world upon which it is added the external world, but rather she is being-in-the-world from the very beginning[9]. Embodiment and playfulness become the key ingredients thanks to which this being-in-the-world, or the lifeworld of the infant, are consolidated into a unitary whole. Winnicott's empirical research has been filtered through the prism of phenomenological inquiry to show how this rich material of thought could be transformed into a phenomenologically inspired theory that would account for the early life of the infant, especially for the archaic field of experience.

References

Barbaras, R. (2011). *La vie lacunaire*. Paris: Vrin.

Barbaras, R. (2021). *Introduction to a Phenomenology of Life*. Bloomington: Indiana University Press.

Fazakas, I., & Gozé, T. (2020). The Promise of the World: Towards a Transcendental History of Trust. *Human Studies* 36, 69–189.

Fink, E. (2016). *Play as Symbol of the World*. Bloomington and Indianapolis: Indiana University Press.

Gallagher, S. (2020). *Action and Interaction*. Oxford: Oxford University Press.

Heidegger, M. (1962). *Being and Time*. Oxford and Cambridge: Blackwell.

Henry, M. (1973). *The Essence of Manifestation*. The Hague: Martinus Nijhoff.

Husserl, E. (1973). *Zur Phanomenologie der Intersubjectivitat. Dritter Teil*. Den Haag: Martinus Nijhoff.

Husserl, E. (1989). *Ideas pertaining to a Pure Phenomenology and to a Phenomenological Philosophy. Second Book. Studies in the Phenomenology of Constitution*. Dordrecht: Kluwer Academic Publishers.

Maldiney, H. (1991). *Penser l'homme et la folie*. Grenoble: Millon.

Merleau-Ponty, M. (1968). *The Visible and the Invisible*. Evanston: Northwestern University Press.

Patočka, J. (1998). *Body, Community, Language, World*. Illinois: Open Court.

Patočka, J. (2016). *The Natural World as a Philosophical Problem*. Evanston, Illinois: Northwestern University Press.

Patočka, J. (2022). *The Selected Writings of Jan Patočka. Care for the Soul*. London and New York: Bloomsbury Academic.

Prášek, P. (2021). Is the "Happening Subjectivity" still a Subject? Marc Richir's Conception of Subjective Identity. *Meta: Research in Hermeneutics, Phenomenology, and Practical Philosophy* 13 (1), 58–82.

Richir, M. (1987). *Phénomènes, temps et êtres*. Grenoble: Millon.

Richir, M. (2004). *Phantasia, Imagination, Affectivité*. Grenoble: Millon.

Richir, M. (2006). *Fragments phénoménologiques sur le temps et l'espace*. Grenoble: Millon.

Richir, M. (2010). *Variations sur le sublime et le soi*. Grenoble: Millon.

Richir, M. (2015). *L'écart et le rien*. Grenoble: Millon.

Richir, M. (2016). *Propositions buissonnières*. Grenoble: Millon.

[9]This last statement points to the idea that the infant does not create the world solely from his/her mind, but rather, he/she discloses it through interactions and encounters with the others.

Ricoeur, P. (1966). *Freedom and Nature: The Voluntary and the Involuntary*. Evanston: Northwestern University Press.

Sartre, J.-P. (2003). *Being and Nothingness*. London and New York: Routledge.

Schutz, A. (1976). *Collected Papers II. Studies in Social Theory*. The Hague: Martinus Nijhoff.

Sheets-Johnstone, M. (2006). Essential clarifications of "self-affection" and Husserl's "sphere of ownness": First steps toward a pure phenomenology of (human) nature. *Continental Philosophy Review* 39, 361-391.

Stern, D. N. (1998). *The Interpersonal World of the Infant*. London: Karnac.

Taipale, J. (2021). Being Carried Away. Fink and Winnicott on the Locus of Playing. *Journal of Phenomenological Psychology* 52, 193–217.

Winnicott, D. W. (1965). *The Maturational Processes and the Facilitating Environment*. London and New York: Karnac.

Winnicott, D. W. (1989). *Psycho-Analytic Explorations*. London and New York: Routledge.

Winnicott, D. W. (2009). *Playing and Reality*. London and New York: Routledge.

Wynn, F. (1997). The Embodied Chiasmic Relationship between Mother and Infant. *Human Studies* 19, 253–270.

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 223 – 238

HOW ARE LARGE LANGUAGE MODELS FALLIBLE: EPISTEMIC PROSPECTS FOR NEURAL INTELLIGENCE

Boris Grozdanoff*

Abstract: Large Language Models (LLMs) operate within closed neural architectures, isolated from direct physical interaction with the external world. This lack of empirical grounding is a critical epistemic deficiency of empirical theories to explain the referential nature of LLM outputs and their hallucinatory fallibility. To address this, I propose a computational adaptation of Fregean semantics that integrates the semantics of transformer architectures with classical philosophy of language. In it I reject static token embeddings as the sole bearers of meaning in an LLM and I formally identify shifted, contextualized vectors generated via self-attention as rigorous computational analogues of Fregean sense. This structure provides an epistemological grounding for how transformers construct meaning, and then it demonstrates the new potential of ungrounded models to execute *a priori* revisions of deeply learned, highly probable "beliefs." I integrate these contextualized Fregean senses into a Quinean-style web of beliefs,[1] endowing the neural network with dual epistemic fallibilities: empirical and *a priori*. To demonstrate this mechanism, I model a paradigmatic historical shift, the a priori revision of the absolute simultaneity principle, as it would unfold within this modified Fregean vector space. I identify a multi-sense consistency mechanism which can correct errors a priori and I show that this approach exhibits a robust model capable of delivering epistemic success in artificial systems devoid of real-world grounding.

Keywords: Large Language Models, Neural Intelligence, Fregean sense, a priori revisability, transformers.

The explosive success of modern transformer-based[2] Large Language Models (LLMs)[3] has initiated an intense interdisciplinary debate on whether artificial neural

[1] I use the term "belief" here operationally and not anthropomorphically. In the context of LLMs, an artificial "belief" does not denote a conscious propositional attitude. Rather, it denotes a stable (after training) highly probable configuration of contextualized embeddings.

[2] Transformers are a specific, novel class of artificial neural networks with a specialized internal structure (heads, keys, queries, values) and a particular mechanism (attention) which brought about the immense success of modern LLMs. See the original paper in Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. [2017] "Attention is all you need". Advances in Neural Information Processing Systems, 30. Accessed at https://arxiv.org/pdf/1706.03762

[3] For a recent survey on LLMs success see Zhao, W. X., et al. (2026). "A survey of large language models". In Frontiers of Computer Science, 20(12), 2012627. https://doi.org/10.1007/s11704-026-60308-3

*Institute for Philosophy and Sociology, Bulgarian Academy of Sciences; grozdanoff@gmail.com

©2026 Boris Grozdanoff. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

networks do have genuine understanding of natural language or just operate as complex "stochastic parrots".[4] LLMs not only generate exceptionally fluent and semantically adequate textual outputs, but they also consistently answer difficult and complex questions correctly. The main and still unsolved problem with their output are the notorious *confident hallucinations*,[5] where an LLM asserts (typically) a sentential output P, which still has the best generative probability, but it is factually false, logically contradictory or both. Hallucinations are an unresolved problem.

The main reason for this vulnerability is the so-called "grounding gap":[6] the models are trained on static text corpora without any direct sensory access to the physical world. In contrast, with human cognition, the most successful epistemology is one or another version of modern empiricism[7]. The empirical verification of all sentences, accepted as fallible in principle ever since Pierce and Quine,[8] serves as the main epistemic correction mechanism: when our epistemic beliefs clash with our physical observations, we revise our mental models of the sentential part of the referred to reality. For ungrounded computational systems, like the LLMs, however, this empirical route of revising fallible sentences is not available. If LLMs are to achieve genuine epistemic reliability and cognitive autonomy, they must possess an internal to their systems mechanism to self-identify, comprehend and correct all categories of errors.

Those errors come in two general categories, *internal to the model*, that is, in cases with Ps which do not require direct access to the physical world, but also, *external to the model*, as in cases with Ps, which do require such access. The mechanism for the correction of the first kind is, I argue, *a priori*, and the mechanism for the second kind is empirical. A mechanism of the second kind can be developed only in systems with sensory "embodiment", such as robotic systems with sensors that deliver flow of empirical data to the system and is not a focus of this paper. But a mechanism of the first kind can be developed without any embodiment of the system. For this purpose, I propose a high-level formulation of a framework where an empirically isolated neural system can optimize its internal representational space, and, as a consequence, its sentential output, through pure reasoning, which is neither triggered by nor executed in virtue of empirical data. The framework combines Gottlob Frege's influential *Sense and Reference model* of meaning with the high-dimensional vector spaces of the transformer neural networks in the LLMs and proposes that we can

[4]The term "stochastic parrots" was used by Bender et al., in their seminal 2021 paper "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" (Bender et al., 2021). In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery.

[5]Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. [2023]. *Survey of hallucination in natural language generation*. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730

[6]Spangher, A., et al. (2025). "NewsInterview: a Dataset and a Playground to Evaluate LLMs' Grounding Gap via Informational Interviews." In Proceedings of the 57th Annual Meeting of ACL.

[7]Brown, James [1993] *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*, Routledge, NY.

[8]Peirce, C. S. [1868] "Some Consequences of Four Incapacities" (reprinted in The Essential Peirce, Vol. 1). pp. 140–157, accessed at https://ia800806.us.archive.org/26/items/jstor-25665649/25665649.pdf

develop an autonomous self-correction module that executes *a priori* revisions within LLMs.

1. Fallibility, Quinean centrality and a priori revisability

In order to construct a robust model of non-empirical revision in artificial LLM systems it is useful to shortly trace how fallibilism and revisions have been addressed in analytic epistemology, from Charles Peirce, through Willard Quine and Michael Friedman, to a recent account by Boris Grozdanoff of *a priori* revisability.

LLM's hallucinations, where the output generates an "invented" answer, nevertheless follow the *exact same model* of generation of "non-invented" outputs, where a sequential next word prediction is chosen on the basis of the computed highest probability for this word, thus building up a sentence, P, which we, as human users, interpret as 1. A sentence, 2. As meaningful, 3. As justified, and sometimes as 4. Delivering actual knowledge. The latest versions of the leading LLMs today, such as Opus by Anthropic, Gemini by Google and ChatGPT by Open AI, demonstrate a truly impressive success in all (1–4) in a significant number of cases, while we at the same time observe a noticeable reduction in cases of registered hallucinations. Nevertheless, the outputted Ps remain fallible in a number of ways, most significantly expressed by hallucinations but also by non-hallucinatory output that we, when comparing it with alternative outputs, would rank lower than others in terms of (1–4). To use Frege's well known example of the Morning and Evening Star,[9] a hallucination would be if the model outputs a response to a question of the sorts "Are the Morning Star and the Noon Star the same star?" Any response which acknowledges the existence of some sort of a "Noon Star" in this context would be a hallucination, for example "Yes they are the same", or "No, they are different", etc.

The modern formulation of fallibilism originates with Charles Peirce,[10] who famously argued that no proposition, regardless of its certainty or strength of justification, is permanently *immune to revision*. Peirce stressed the fact that an inquiry is a continuous and self-correcting process, and because our cognitive contact with reality is always mediated by signs, our theories are at best only provisional approximations of an unreachable final reality that can always turn out to differ from even our best theories of it. A productive way to apply the Peircean fallibilism to AI could be to shift from the view of LLMs as static "oracles" of truth-candidates to a view of LLMS as active participants in an ongoing process of inquiry. For our present purposes here those inquiries might be *external*, such as the human labelling of outputs during the training phase (at the RLHF method of enforcing better responses by hand rather than from within the model), or *internal*, where the inquiry is being done entirely within the model.

[9] Frege, Gottlob (1948) "On Sense and Reference" in The Philosophical Review, Vol. 57, Issue 3, pp. 209–230.

[10] Peirce, C.S. [1868] "Some Consequences of Four Incapacities" (reprinted in The Essential Peirce, Vol. 1). pp. 140–157, accessed at https://ia800806.us.archive.org/26/items/jstor-25665649/25665649.pdf

The principle was influentially accepted by the epistemology community and continues to be one of the very rare examples where a significant number of epistemologists agree. Perhaps the most influential historical systematization of fallibility was done by Quine in *Two Dogmas of Empiricism* (1951),[11] where he structured fallibilism in his famous Web of Beliefs by representing our beliefs as nodes in an *interconnected web* impinging on experience *only at its periphery*. Within this web, Quine introduced the dimension of the resistance of any individual belief to revision, which is determined by a new property, introduced in the web structure as "centrality" C(P).

At the *periphery* of the web are observation sentences, which are typically direct sensory reports, but also sentences which contain claims verifiable or refutable via empirical check. The conditions for establishing their fallibility are entirely empirical and they are revised with empirically justified defeating sentences. Thus, if a peripheral example is a sentence like "Venus has a yellow atmosphere", a sufficient set of observational sentences, delivering empirical information about an actual colour of Venus would serve to confirm or reject the truth of this sentence. The truth conditions are quite clear and are in a way "projected externally" from the sentence. They require empirical information that is relevant and adequate to the content of the sentence, otherwise the sentence remains with an unverified claim. It is also clear that if a candidate for defeating or confirming sentence does not contain such information, or does not contain any empirical information at all, it could serve neither as a defeater nor as a confirmer.

At the *core* of the web are very different beliefs, which construct the structural scaffolding of the system, such as the laws of logic, mathematical axioms, and expressions about foundational physical constants. For example, if we take as such a belief, formulated as a logical sentence, *modus ponens* from propositional logic, we are in a very different truth conditional situation. The belief does not make or contain an empirical claim and describes an abstract relationship between a very general class of natural language expressions, formalizable in propositional logic. The only way to defeat it is to demonstrate, with the same logical rigour which makes it logically justified, that the expressed relationship between P and Q does not hold for at least one P or one Q for which the implicatory relationship at the first part of *modus ponens* does hold. This, however, is logically impossible so no logical defeaters exist against it, and this adds to the high centrality of this logical law. Could it be that an empirical belief is able to revise *modus ponens*? For example, if we consider concrete empirical sentences for P or Q? We observe that it cannot, because the claim of *modus ponens* is not an empirical claim, that is, it does not refer to an empirical portion of reality; instead, it is a very specific logical relationship for abstract sentences, and devoid not only of empirical but of any content, the law shapes only their logical relationship. Thus, we are facing a dilemma: on the one hand, Quine fully accepts Pierce's fallibility of all sentences, but on the other, we do not see what method of revision would work for such non-empirical core claims? Should we abandon universal fallibility, or can we preserve it and find a method of revision that can actually examine such core claims? Quine argued that when a peripheral observation conflicts with experience,

[11] Quine, W.V.O. (1951) "Two Dogmas of empiricism" in The Philosophical Review 60 (1951): 20–43.

the resulting tension between clashing beliefs propagates inward and leads to a redistribution of truth values within the web. In order for the system of the web to remain stable, the cognitive agents should follow the "principle of minimum mutilation," where the priority is to change beliefs at the peripheral nodes and not beliefs at the core. However, because Quine's web relies exclusively on the famous "tribunal of experience"[12] to initiate and execute a revision, *it does not provide a mechanism for ungrounded computational systems, like the LLMs, to revise core beliefs*. The lack of a specific mechanism that would allow one to identify the need for and execute revisions of core, non-empirical beliefs, was historically pressing after Quine's leading role of empirical revisions. Perhaps the most influential attempt in the direction to fill his vacuum was Michael Friedman's dynamical model of scientific knowledge. In *Dynamics of Reason* (2001),[13] Friedman addresses the limitations of scientific knowledge and in particular its revision dynamics, by proposing a three-layered model. The structure has a base empirical layer of physical laws, a philosophical meta-paradigm layer of meta-theoretical guidelines, and a key, novel, second-level *constitutive a priori layer*. This layer provides the independent mathematical, spatial, and temporal basis that makes empirical testing possible within a specific paradigm, and effectively it acts as a relativized *a priori framework that has revision dynamics in it*. This model, however, does not resolve the problem with revision of non-empirical beliefs, as it cannot provide a purely epistemic explanation of how these constitutive frameworks are revised during conceptual revolutions without ultimately collapsing into empirical revisions of the Quinean kind.

Recently Boris Grozdanoff proposed a specific, purely epistemic mechanism to resolve this problem. He formulates a theory of *a priori* revisability,[14] where non-empirical beliefs, which Quine puts at the core of his web, remain fully revisable, in accordance with Pierce's universally accepted fallibility, and at the same time, are *non-empirically revisable*, contra Quine. Their non-empirical, a priori epistemic character is determined in a novel, two-dimensional way: non-empirical beliefs might be a priori justified, that is, the nature of their justification could be a priori, but they also could be non-empirically revisable, that is, a priori revisable. Both dimensions add significant epistemic strength and structure to the traditionally more or less general a priori tenet for general and fundamental logical, mathematical and physical-constitutive beliefs, which Quine would put at the core of the Web.

Grozdanoff argues that core structural principles are truly independent of experience and thus fully *a priori* only when they are independent of experience across both dimensions. Thus, he allows for two distinct a priori horns of an allegedly a priori belief: the justificatory horn, where the reasons to hold that P are independent from experience, and the revision horn, where the reasons to revise that P are independent from experience. To establish this, he formulates two independent epistemic criteria:

[12]Quine, W.V.O. [1951] "Two Dogmas of empiricism" in *The Philosophical Review* 60 (1951): 20–43. See Part V, The Verification theory and Reductionism for "tribunal of sense experience".

[13]Friedman, Michael (2001). *Dynamics of Reason: The 1999 Kant lectures*, CSLI Publications, Stanford, p. 71.

[14]Grozdanoff, Boris [2014] *A Priori Revisability in Science*, Cambridge Scholars Publishing.

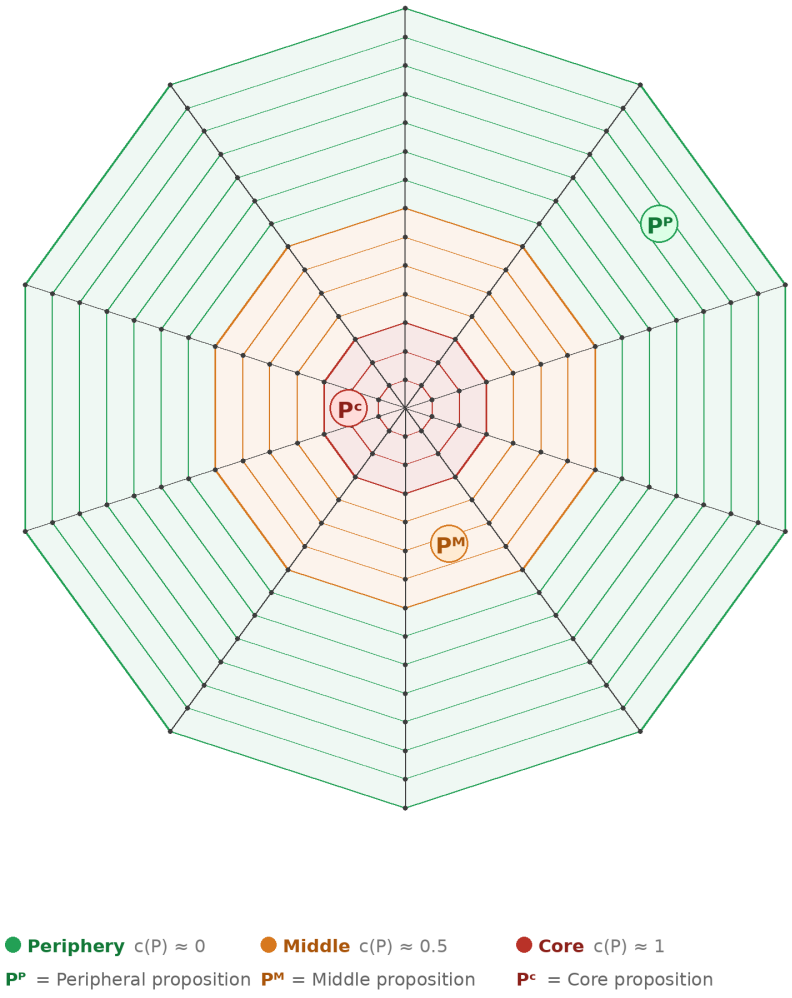


Figure 1. Quinean style Web of beliefs, Beliefs are denoted as nodes

Justificatory Independence: A proposition is *a priori* justified if its warrant does not rest on sensory evidence.

Revisability Independence: A proposition is *a priori* revisable if the reasons for abandoning, modifying, or replacing it are entirely non-empirical.

Grozdanoff's dynamical epistemic framework is rationalistic and contains two different epistemic streams of revision dynamics. The *a priori stream allows for a non-empirical belief to be revised* because it is found to cause an internal, logical inconsistency or a structural tension within the system of beliefs. This "defeating procedure" preserves fallibility but at the same time allows non-empirical beliefs to be revised in the most adequate epistemic way: non-empirical beliefs, in virtue of (most of) them not having empirical claims, besides the possibility to have non-empirical

grounds to be held *should also* have the same type of grounds not to be held, that is, to be revised.

This model, even if developed for the purposes of scientific epistemology and not computer science, is remarkably applicable for the fallibility needs of modern LLMs. Since LLMs are not grounded and lack physical world sensors, they cannot perform empirical revisions; however, one thing they can do very well is to identify structural and logical conflicts within their representational spaces and thus perform non-empirical, *a priori* revisions.

2. Artificial Fregean sense structure in transformers

To implement the dynamics of a priori revisability in artificial systems, like LLMs, we can establish a structural mapping between Gottlob Frege's classic semantic theory of sense and reference and the architecture of modern Transformers. In his seminal work *On Sense and Reference* (1892), Frege famously resolved the paradox of how identity statements like $a = a$ and $a = b$ differ in cognitive significance by dividing meaning into two dimensions: *Reference*, the objective entity or truth value a sign designates, and *Sense*, the "mode of presentation" of the reference. Frege observed that multiple different senses can point to a single reference, and an expression can have a meaningful sense even if it does not have a referent.[15]

Alternative semantic accounts, like the Davidsonian truth-conditional semantics or Brandom's Inferentialism develop robust and influential models of meaning in the analytic philosophy of language; however, the Fregean framework is uniquely suited for implementation in transformer based neural architectures. Transformers natively exhibit a one-to-many relationship between a single vocabulary token and its potentially infinite number of contextualized embeddings (the Fregean senses). Even historically the Fregean framework was designed to solve an isomorphically identical semantic puzzle, namely, how identical referents generate different cognitive significances. That is why I believe this choice is superior to alternative semantic accounts and it offers a geometrically compatible structural mapping for the self-attention mechanism in transformers.

This Fregean semantic structure can be mapped onto the vector space of a transformer neural network. To structure this we can use the core Fregean Sense and Reference, split between the *Word Level* and the *Sentence Level*.

In order to understand how to map Fregean structure over the neural architecture of LLMs we need to have a high-level understanding of how a transformer neural network processes natural language input as words and sentences. At input, when some text enters a transformer, the model translates it into a very structured, multi-dimensional geometric universe where the text begins to live as vectors in n-dimensional vector space, called *latent space*. The number n defines the number of the dimensions of the space and it does not depend on, say, the training text dataset

[15] Frege, Gottlob [1948] "On Sense and Reference" in *The Philosophical Review*, Vol. 57, Issue 3, pp. 209–230.

volume, but is an architectural choice of the system architect and usually has 1024 or even 4096 dimensions.[16] In this space, vectors are the only elements, but unlike pure geometry, in computer science vectors are not made up of geometrical points but of numbers and those numbers are coordinates. Every word and sentence exists only as an array of numbers. The word "Star" is a vector, and its numerical values, say [0.90,0.00,0.85][17], are coordinates.

Here it is appropriate to acknowledge that the static word embeddings in LLMs are inherently products of distributional semantics operating according to the principle that meaning is to be derived from statistical co-occurrence. However, the distributional semantics explains fully the initial generation of the static and non-contextualized geometries, but *it cannot explain the dynamic fixing of the reference and the truth value assessment* which occurs at the sentential level. The Fregean framework delivers the structure which fills in the semantic (in terms of meaning and reference) and epistemics abyss between the word-associations and the compositional sentence-senses.

To carve a Fregean sense in this space, using his example with *The Morning Star = The Evening Star*, when in fact they both refer to the planet Venus, we need to choose axes that can capture the actual properties of the cosmic entity in question along our descriptions of it. Let us consider a very simple 3-d space, where the three axes represent the concepts of [Planetary Identity, Dawn Alignment, Dusk Alignment]. After the model is trained on a text corpus, it "learns" a map of static coordinates for the words with respect to these (and in our case – only these) particular conceptual dimensions:

The word "Star" might have static coordinates [0.10, 0.00, 0.00]. Standing completely alone, the word "Star" has a very low score on the planetary axis 0.10 (because stars are stars and not planets) and is entirely neutral (0.00) to specific times of day like *dawn* or *dusk*.

The word "Morning" might have coordinates [0.00, 0.95, -0.95]. It has nothing to do with planets (0.00), but aligns very much with dawn (0.95) and also points in the exact opposite direction of the dusk axis (-0.95).

The word "Evening" might have coordinates [0.00, -0.95, 0.95]. It is also a non-planet (0.00) but it also points opposite to dawn (-0.95) and very much aligns with dusk (0.95).

The concept of "Venus" lives in the map as a fixed vector with coordinates [0.95, 0.50, 0.50]. It is a planet (0.95), and because it orbits closer to the Sun than Earth, as the trained model already knows, it is visible from Earth both at dawn and at dusk (say, 0.50 on both time axes).

[16] Chen, Y. et al. [2025]. "Not All Hidden-States' Dimensions are Needed in Transformer", In GitHub/MLResearch. Accessible at https://raw.githubusercontent.com/mlresearch/v267/main/assets/chen25bc/chen25bc.pdf

[17] All given coordinates, both "learned" and after attention operations values, are mock values and serve only as illustration. Actual values are multidimensional, and in reality, it would be enormously difficult for these to be extracted and used as illustrations of a model.

When these words build up a prompt they do not remain at their static coordinates anymore. The self-attention mechanism of the transformer network makes them operate and moves the vectors to new locations. For example, when the network processes "The Morning Star" its attention heads mathematically "blend" the coordinates of "Morning" into "Star." The generic "Star" vector is moved from its "learned" location and its planet score increases because the model now recognizes the astronomical context, and its dawn score increases too. Its new, *contextualized* coordinate shifts to [0.75, 0.85, -0.30]. *This shifted coordinate functions as a computational analogue of Frege's Sense for the Morning Star.* The traditional Fregean sense carries a cognitive significance for the human-cognitive agent and here we observe that the vector shift carries a computational significance for the artificial agent.[18] The new, altered geometric position, determines uniquely the further attention pathways and probability distributions in the neural layers to follow. In this way the model "grasps" the reference *differently* due to the intensional vector shift caused by the context.

The main additions of endowing the LLM vector space with Fregean sense are twofold: first, there is a semantic addition, which analyses the meaning of sentences in a productive way and second, there is the epistemic since it would allow us to semantically drive the revisions of the LLM's "beliefs". When the network processes "The Evening Star", another attention operation occurs in the same sentence. The second "Star" vector is moved toward "Evening", and it shifts its coordinate to [0.75, -0.30, 0.85]. *This shifted coordinate is the Sense for the Evening Star.*

So far, the model produced two different vector coordinates, which we can map to two Fregean word-senses for the word "Star" because they figure in two different contexts. Now the system must build up, out of the word vectors, a sentential vector for our sentence "The Morning Star is the Evening Star." *In the deeper layers of the neural network the transformer executes an operation called "pooling" where the system mathematically fuses the dynamic coordinates of the Morning Star vector and the Evening Star vector along with the relational vector for the word "is."* The result of this geometric synthesis is a single new coordinate, in the form of *a single sentence-embedding vector.* This vector is the mapping of the *Fregean sense for the sentence.*

We see that in this space *the direction and the distance represent meaning,* and specific axes represent conceptual features such as astronomical properties or any other property learned from the training corpora, for that matter. Words are not static definitions in a dictionary but are natural language elements quantified into numerical entities, which inhabit the latent space. Within this structure, every concept is defined by its relative position to every other concept.

The remaining word-reference, for both the Morning star and the Evening Star expressions, is anchored in the model in an entity with the name Venus, which has a sense of its own, and it is a planet, not a star, which rises in the morning (like the

[18]Which, for the time being, I will refrain from calling an "artificial cognitive agent", the artificial cognition which approximates human cognition, should have a much greater complexity than the one we see in LLMs along with some elements of it are neither simulated nor integrated in an AI model I know of.

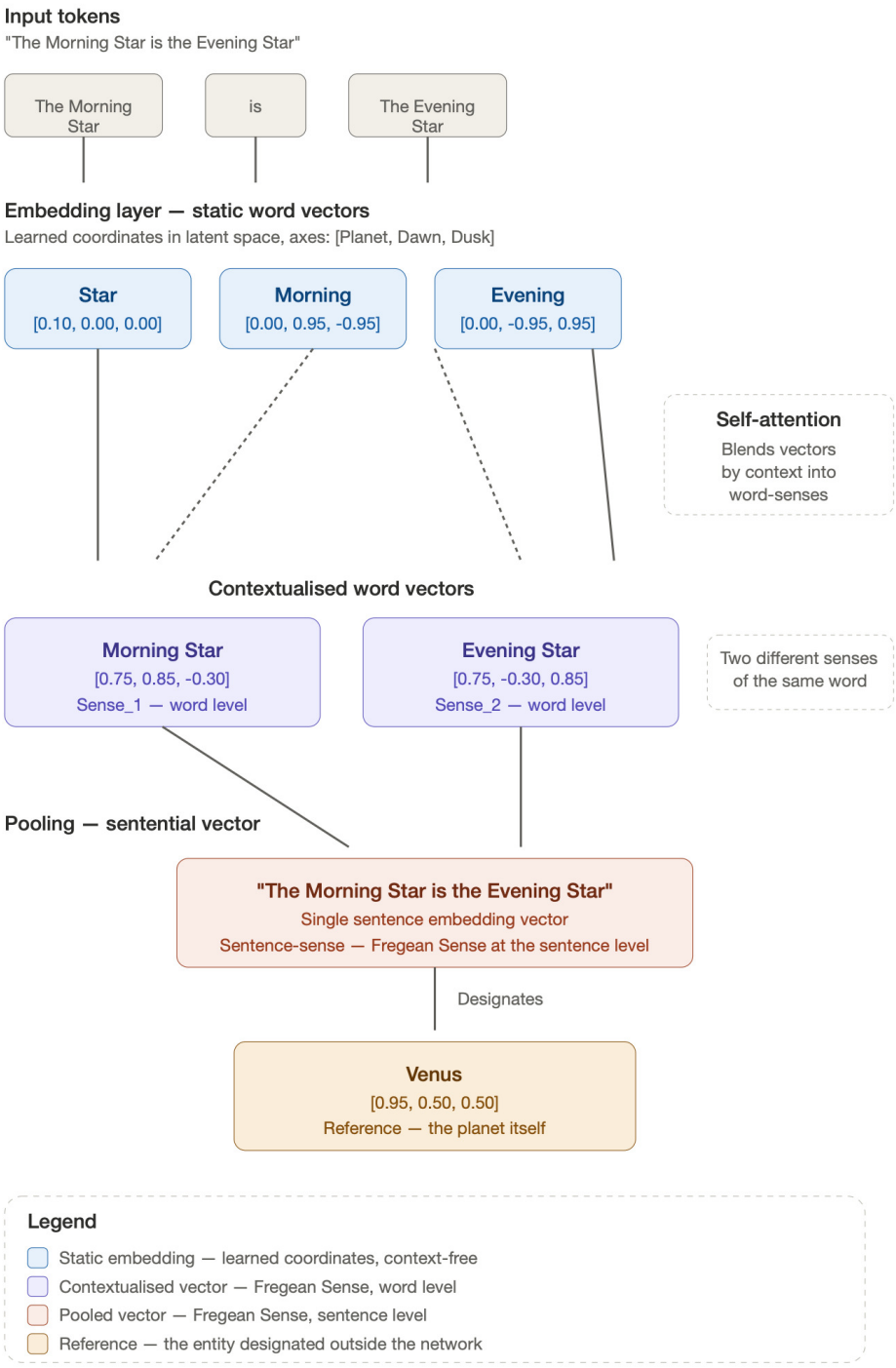


Figure 2. Fig. 2: Fregean Senses mapped on transformer's vector space

Morning star) and sets in the evening (like the Evening Star, etc.).[19] The vector for

[19]Venus has been a familiar cosmic body since at least the Babylonians, but for modern observations with telescopes see research on Galileo's observations in Palmieri, P. [2001] "Galileo and the discovery

Venus is in the model, but when outputted in a sentence, like "The morning Star is the Evening Star and it is the planet Venus" it is artificially "uttered" as a statement and has produced the necessary sense formulation and manifestation that would allow for a cognitive agent, be it the model or a human, to identify and observe the reference in the physical world. A human might observe Venus via a telescope, and an embodied LLM system might do the same. But in reality, the reference of the artificial sentence is (in this case) outside the model itself, just as Venus is outside a human uttering of the same sentence.

3. Artificial a priori revisability in transformers

The model of a priori revisability we will implement here is the one developed by Grozdanoff (2014).[20] While the predominant interpretations often interpret Einstein's formulation of the Special Theory as empirical, Grozdanoff formulates a contrary epistemic account based on historical analysis that at least one of the key principles of STR was arrived at and justified non-empirically. He argues that Einstein's famous revision of the Principle of Absolute Simultaneity was conducted *fully a priori*, as the epistemic nature of the defeaters was not old or new empirical data, which he argues, was not only unavailable at the time, but, even worse, all accessible empirical data was contra the principle of relative simultaneity and in support of Absolute simultaneity. Grozdanoff reconstructs the defeater, Einstein's Train Though Experiment, as an a priori revision justified by internal logical and scientific-framework structural tension. Notably, the epistemic nature of the defeater is not provided by new empirical data. What is even more notable, contra Quine's ultimate Tribunal of Experience, the fallible principle, which got revised, the Absolute Simultaneity principle was *empirically* justified. So, the power of a priori revision was historically demonstrated to hold not only over a priori justified principles, which would be milder and perhaps more acceptable for an empiricist, but over some empirically justified beliefs as well.

Historically the revision of absolute simultaneity was not caused by observational anomalies, say, a team of physicists measuring picosecond temporal differences in the simultaneity of lightning strikes Front and Back as viewed from the train and from the embankment in the thought experiment, but by the discovery of a strong structural contradiction within the Newtonian "Web of Beliefs", which was universally accepted before 1905. The contradiction was observed by Einstein when he attempted to uphold the Principle of Relativity (PR) and the Light Principle (LP) *together*. In the Train Thought experiment, Einstein demonstrated that maintaining LP together with PR both mathematically and logically rejects time as an absolute and universal coordinate structure for all reference frames.[21] The tension was unsustainable for

of the phases of Venus". In *Journal for the History of Astronomy*, 32(2), 109–129. https://doi.org/10.1177/002182860103200202

[20] Grozdanoff, Boris [2014] *A Priori Revisability in Science*, Cambridge Scholars Publishing. See Ch. 4.

[21] Einstein's Train thought experiment is formulated in Einstein, A. (1916) *Relativity: The Special and the General Theory*, Methuen & Co Ltd., in Chapter 9, titled "The Relativity of Simultaneity", pages 25–27.

physics and this forced the a priori revision, which rejected the empirically justified absolute simultaneity and substituted it, in an a priori way, with an a priori justified principle, the Relative simultaneity principle. Its justification was purely a priori, not only in a mild sense that was merely independent of experience, as it was contradicted by experience, but rather the strongest form of "independence" there could be. The revision eliminated the tension and restored the logical coherence of the theoretical framework and shows that the core in Quinean sense principles can be rationally revised from within the web without requiring direct empirical grounding. Modern LLMs exhibit remarkable isomorphism to a Quinean-like web of beliefs, albeit with the proviso that they are structured with our artificial Fregean sense and reference in the transformer systems.

To see how an LLM could autonomously perform a priori revisions of LLM sentences, grasped as shifted sentential vectors, representing a Fregean kind of sentential sense, we can mock trace how a hypothetical Einsteinian-like revision of the Absolute Simultaneity principle would look in our Fregean enhanced transformer vector space. This would allow us to identify and formulate an artificial a priori revisability mechanism so that we can then encode it as a separate module to be added to open source LLMs for testing.[22]

Let us now examine how an a priori revision would look in the internal structure of a transformer network enhanced with our Fregean semantic mapping. The model's vector space could be regarded as a Quinean Web of Belief, with centrality and fallibility of artificial beliefs, yet the revisions allowed here are both empirical and a priori, unlike the original Quinean web. The static word embeddings of the natural language words in a trained model are uncontextualized; the contextualized, shifted vectors deliver Fregean word-senses and the pooled synthesized vectors deliver Fregean sentence-sense. All references (like physical trains, observers, platforms, planets, etc.) of words and sentences, remain fully outside the model. In this way any revision internal to the model that it performs must be *structurally a priori*; either in a *strong a priori* sense driven by the detection and resolution of internal structural contradictions, or a *weak a priori* sense, when ultimately revision is traced to empirical data in the model – for instance, if an empirically represented tension in the model triggers the revision. The static embeddings of the model are acquired via (indirect) empirical exposure to training corpora, but the mechanism of the revision itself operates entirely independently of *new empirical data*. Following Grozdanoff's criterion of revisability independence, the revision is driven purely by the geometric and logical resolution of the internal to the model semantic contradictions during the self-attention phase. This effectively renders the process of correction functionally and structurally a priori.

Let us imagine a "Newtonian" LLM. The model is exclusively trained on classical physics and pre-special theory of relativity electromagnetic theory. Its space contains

[22]The code development and the implementation of this module is beyond the present goals of this paper. However, given the structure offered here, in my view it could be easily developed, integrated and tested within an open-source transformer based LLM system.

three specific conceptual axes to evaluate semantically, using Fregean senses the revision of absolute simultaneity: (Absolute Universal Time, Invariance of Physical Laws,[23] Constancy of Light Speed[24]).

In its initial state, before a prompt, the model has static word vectors with learned coordinates which express the learned meaning of a word, as per the training corpora. The word "Simultaneity" is a static vector with coordinates very much weighted toward classical mechanics, say [0.95, 0.10, 0.00], it is very heavy on the absolute universal time (0.95), it is weak with respect to the invariance of laws (0.10), and is neutral to light speed (0.00). The word "Laws" (here representing the Principle of Relativity) is, say, at [0.00, 0.95, 0.00], and highly towards the invariance. The word "Light" (representing the Light Principle) is, say, at [0.00, 0.00, 0.95]. At this static stage the vectors live together with no tension at all because they are just isolated and learned dictionary definitions. They don't make claims yet.

Now, let us introduce Einstein's Train thought experiment in the context windows of the model: Two bolts of lightning strike both ends of a moving train, front and back, simultaneously as seen by an observer on the embankment. The speed of light is constant. What does the observer on the train (standing at the exact middle of the train!) measure about the simultaneity of the two lightning strikes, at the exact instant when the two observers' positions coincide?

The words and sentences of the text enter the input layer of the transformer and begin to propagate through the neural layers, the self-attention mechanism begins to operate, and blends the static vectors mathematically into contextualized Fregean word-senses with shifted, novel positions in the vector space. The attention heads of the transformer calculate the dependencies between the words "Light," "Moving Train," and "Simultaneous." Now, *the model must impose its deeply held structural principles (PR and LP) onto the new context, because they have the greatest probability power for predicting the output given learned locations of the static vectors and the context of the prompt.* For the platform observer, the light from both strikes travels at constant speed c over same distances and arrives at the same time. The word-sense for "Simultaneity (Platform)" forms very naturally *without any tension* with any other formed "artificial belief" in the vector space, keeps its very high absolute time coordinate value, shifts a bit to [0.90, 0.40, 0.85] as it needs to harmonize with the light principle within the prompted context.

However, when the attention heads try to construct the word-sense for "Simultaneity (Train)" a logical tension, in virtue of clashing semantic senses, takes place! The semantic and the logic of the prompt, in the model now governed by the most heavily weighted vectors of PR and LP, *infers that* the train observer is moving toward one light beam and away from the other. Because the speed of light must remain constant (keeps a value of 0.95 on the LP axis), the geometric projection of the light vectors *requires that the beams arrive at different times for the train observer.*

[23]Einstein, A. [1916] *Relativity: The Special and the General Theory*, Methuen & Co Ltd., in Chapter 9, titled "The Relativity of Simultaneity", For the principle of Relativity see pages 15–17.

[24]Einstein, A. [1916] *Relativity: The Special and the General Theory*, Methuen & Co Ltd., in Chapter 9, titled "The Relativity of Simultaneity", For the Light Principle see pages 19–21.

The logical and semantic tension now takes the form of an outright logical clash: the learned sense of "Simultaneity" absolutely requires a single time value for all frames, which would represent the absolute simultaneity principle; as absolute simultaneity could only have 1 value of occurring events, in our case a simple version of only 2 events, and would anchor its vector to the [0.95] coordinate on the Absolute Universal Time axis.

In the transformer's space this, of course, is not represented as a physical impossibility but as a grave mathematical and logical tension which the system cannot live with. The dot product operation between the contextual requirements of the Light Principle and the classical constraints of Absolute Simultaneity gives a high negative value. The vectors are moving in orthogonal and clashing directions. If the network forces these contradictory word-senses towards the final pooling layer in the transformer, the resulting sentence-sense vector would simply collapse into high entropy and semantic incoherence. Effectively, *the model has discovered a logical contradiction purely a priori*: the Web of Belief of the model cannot simultaneously maintain PR, LP, and Absolute Simultaneity given (1) the vector geometry of the thought experiment and (2) the vector geometry of its learned static vectors, all of them shifting in the space towards a coherent sentential sense output with neither logical nor semantic tension.

Now, there are only two scenarios. One scenario is that a standard model which lacks an explicit structural revisability mechanism might hallucinate an answer based on a priori probabilities. Or, alternatively, new SOTA models can attempt to resolve the tension by a sequential "Chain-Of-Thought" (CoT) reasoning. However, CoT fundamentally remains a probabilistic generation of intermediate tokens. It lacks a formal, geometry-driven semantic inconsistency penalty which has been designed to identify and logically resolve structural clashes a priori. In practice, a standard model with no a priori revisability mechanism module integrated, would simply freeze or hallucinate one wrong answer or another, dug up from the vast pool of trained highly probable options it has in its bottomless static vector weight parameters (recently, some large language models have exceeded 1 trillion parameters). Most probably, the outputted answer would be some hallucination with a positive probability of the output, fully ignoring the mathematical, the semantic and the logical clashes. Or, second, if the model figures an integrated a priori revisability mechanism, it would execute an a priori revision that would fully resolve all three axes of tensions, and it would thus revise the absolute simultaneity core principle and substitute it with the relative simultaneity principle. The mechanism would have resolved the otherwise unsurmountable architectural tension in the space before the selected output, while, at the same time, it would have discovered one of the leading principles of the Special Theory of Relativity.

4. A priori revisability module for transformers

When we successfully integrate a transformer mechanism for a priori revisability, we effectively elevate *the network from a probabilistic engine to a rational epistemic agent which operates over a Fregean semantic geometry*. In standard models next word predictions are delivered, almost exclusively, by harnessing probabilities. The current enormous success of LLMs notwithstanding, output is based on highly repeated concepts, such as the classical principle of absolute simultaneity, being immensely probable, which pulls them toward being predicted in an output with much greater likelihood than infrequent concepts like relative simultaneity in our pre-Newtonian LLM example. The conceptual model proposed here that is enhanced with Fregean word-senses and sentence-senses, modified Quinean fallibility and Grozdanoff a priori revisability should allow, when implemented and tested, the transformer to autonomously bypass statistical determinism when a structural clash is encountered. Functionally, the a priori revisability module would enable the network to continuously monitor its contextualized attention vectors (the word-senses) during the process of them being pooled into sentential embeddings (the sentence-senses). When the model processes upon a prompt a text where conceptual vectors that are deeply "learned" after training conflict logically, semantically and mathematically, the system would be able to detect this clash. To resolve this without relying on external empirical grounding, the system would rank the structural importance of the conflicting beliefs (embedded as shifted sentential vectors). Then, it would compute a *composite structural penalty* against the standard probability of the generated output and, depending on the computed result, it would be able to override and reject highly probable but logically clashing claims in favour of statistically improbable albeit geometrically, logically and semantically coherent outputs. Technically, within the neural model, this contradiction detection mechanism would compute the cosine similarity between the projection of the newly generated sentential sense vector and the learned geometrical hyperplanes of the core Quinean style principles. If the dot product renders a strong negative value, which would indicate clashing, orthogonal directions, the mechanism would apply the penalty to the standard generation probabilities. Depending on the computed result, the penalty would enable the system to forcefully ignore highly probable but clashing claims, and thus it would direct the output towards geometrically non-clashing outputs.

The paper developed a conceptual framework which synthesizes the Freegan sense and reference semantics, Quinean epistemology and the high dimensional vector geometry of the neural transformer architectures to address the epistemic grounding gap in Large Language models. By mapping contextualized embeddings as computational analogues of Fregean sense and by structuring them with a dynamically revisable, modified, Web of Beliefs, the proposed model demonstrates how ungrounded neural systems can theoretically execute a priori revisions of logically contradictory yet highly probable generations. The proposal puts forward a programmatic philosophical foundation for autonomous error correction as a proof-of-concept and not as a fully deployed algorithmic solution. The validation of the a priori revisability module requires the transition from analytic epistemology to empirical computer science. A future research must formalize the functions which will govern the structural penalty,

must establish a rigorous semantic inconsistency metrics to evaluate geometric clashes via cosine similarity, and must develop specific hallucination reduction benchmarks to test the module's practical implementation, most probably within accessible open source LLM models.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).

Brown, J. R. (1993). *The laboratory of the mind: Thought experiments in the natural sciences*. Routledge.

Chen, Y., et al. (2025). *Not all hidden-states' dimensions are needed in transformer*. MLResearch. https://raw.githubusercontent.com/mlresearch/v267/main/assets/chen25bc/chen25bc.pdf

Einstein, A. (1916). *Relativity: The special and the general theory*. Methuen & Co. Ltd.

Frege, G. (1948). On sense and reference. *The Philosophical Review, 57*(3), 209–230.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).

Brown, J. R. (1993). *The laboratory of the mind: Thought experiments in the natural sciences*. Routledge.

Chen, Y., et al. (2025). *Not all hidden-states' dimensions are needed in transformer*. MLResearch. https://raw.githubusercontent.com/mlresearch/v267/main/assets/chen25bc/chen25bc.pdf

Einstein, A. (1916). *Relativity: The special and the general theory*. Methuen & Co. Ltd.

Frege, G. (1948). On sense and reference. *The Philosophical Review, 57*(3), 209–230.

Friedman, M. (2001). *Dynamics of reason: The 1999 Kant lectures*. CSLI Publications.

Grozdanoff, B. (2014). *A priori revisability in science*. Cambridge Scholars Publishing.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). *On faithfulness and factuality in abstractive summarization*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Peirce, C. S. (1868/1992). Some consequences of four incapacities. In *The Essential Peirce* (Vol. 1, pp. 140–157). https://ia800806.us.archive.org/26/items/jstor-25665649/25665649.pdf

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review, 60*, 20–43.

Spangher, A., et al. (2025). *NewsInterview: A dataset and a playground to evaluate LLMs' grounding gap via informational interviews. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. In *Advances in Neural Information Processing Systems* (Vol. 30).

Zhao, W. X., et al. (2026). A survey of large language models. *Frontiers of Computer Science, 20*(12), 2012627. https://doi.org/10.1007/s11704-026-60308-3

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 239 – 253

HUMAN TRANSFORMATION IN THE DIGITAL AGE

Victor Mishchenko*

Larisa Gren**

Olga Shymchenko***

Abstract: The article explores the prospects of human transformation under the influence of information technologies. In addition to the traditionally discussed "optimistic approach," which emphasizes the creative and developmental potential of modern technologies, and the "critical approach," which highlights the growing cognitive limitations of individuals, the study introduces a "systemic and realistic approach." This perspective enables a comprehensive evaluation of the nature and causes of personality transformation. The analysis considers several factors: human behavior in the information space (anthropological approach); the influence of simulacra and virtual representations (informational approach); the mechanisms of social control and power (socio-political approach); and the technological transformation of human nature itself (technological approach).

Taking into account the actual power of these influences and current sociological observations – such as the increasing immersion of humanity into the Internet environment, the shift in cultural values, and the changing interests of younger generations – the article identifies the main trend in shaping the human being of the future. It traces the evolution from the individual of the information society to the human immersed in virtual space, and ultimately to the posthuman of the emerging nanotechnological era.

Keywords: human transformation, information technology, virtual reality, information society, simulacra, cognitive change, digital culture, technological evolution, socio-political control.

1. Introduction

Our world of the future, which we are creating through the efforts of generations living and existing today on planet Earth, depends on our economic power and capacities, our ideals, our understanding of beauty, harmony, and environmental sustainability. Yet, in addition to our economic capacities and spiritual aspirations, the world of the future also depends on what kind of human being will inhabit this new world.

Through centuries, the human being has undergone change. History recognizes the human of the ancient world, who sought harmony with the cosmos, nature, and the laws of the polis. It recognizes the human who discovered the mystery of communion

*Department of Philosophy, Kharkiv Polytechnic Institute, Ukraine; ORCID: 0000-0002-8815-9272; viktorim2017@gmail.com **Department of Pedagogy and Psychology of Management of Social Systems named after Ivan Zyazun, Kharkiv Polytechnic Institute, Ukraine; ORCID: 0000-0003-4466-6018 ***Department of Philosophy, Kharkiv Polytechnic Institute, Ukraine; ORCID: 0009-0001-8514-5596

©2026 Mishchenko, Gren, and Shymchenko. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

with higher forces. The human who experienced the joy of life in self-development and self-realization. The human who created a gigantic technosphere that transformed the appearance of the planet. However, despite this diversity, since the Axial Age, when, according to Karl Jaspers, a human being close to our modern type of personality was formed, "the foundation of modern humanity" was established.

Yet, over the past several decades, the world has begun to change rapidly. The transformation of the modern human being is taking place under the influence of information technologies. The pace of this transformation is so great that parents no longer understand their children, while children of different age groups already speak different languages and embrace different values. What image will the human of the twenty-first century assume? What factors influence its formation? The resolution of these questions determines not only what society will become, but whether it will exist as a society in the traditional sense – with its collective will and aspiration toward collective creation – or as a conglomerate of individuals, each dissolved in the virtual space of the world of information technologies.

Humanity must be prepared for the threat of a roboticized future, in order to comprehend the nature of the transformational process and to prevent negative tendencies. Therefore, it is necessary to concentrate on the image of the new human being, who is under the influence of information technologies and who is already visibly different from previous generations in the very character of their existence in the world. How can one assess the impact of such technologies on the human being?

It is necessary to identify the general trend, the nodal points along the path of human transformation in the digital society, to recognize the overarching logic of the transformation of personality depending on the degree of influence of the informational and technological factors of modern civilization.

2. Theoretical basis

In the vast array of information regarding the influence of the digital society on the individual, it is possible to distinguish optimistic, critical, realistic, and analytical contexts, which reveal different forms of influence (see Table 1).

Approaches that analyze the consequences of exposure	Approaches that analyze the causes of impact
Optimistic	Anthropological
Critical	Informative
Realistic (or comprehensive,) systemic	Socio-political
	Technological

Table 1. Main areas of research on information influence.

The formation of the optimistic approach, within which the positive aspects of the influence of digitalization processes on the human being are considered, began at the turn of the centuries and was closely connected with the works of the theorists of the information society. The human being significantly expands the volume of information, which provides the possibility for making decisions regarding the optimization

of the processes of managing the market economy. Bill Gates wrote about the ability to implement extraordinary and unique projects (Gates, 1999), Alvin Toffler pointed to the rapid exchange of knowledge (Toffler, 1980), Ray Kurzweil and Terry Grossman to information technologies as an instrument for surpassing the biological qualities of the human being (Kurzweil, Grossman, 2009). Quite consistently, the advantages of information technologies are considered in the works of the French philosopher Pierre Levy, where he defines a new form of scientific cooperation as anthropological unity (Levy, 1994).

Among the adherents of the critical approach is the American scholar and publicist Nicholas Carr, who argues that the development of Internet technologies negatively affects our ability to concentrate and reason and impairs human cognitive abilities, including creativity and the capacity for deep thinking (Carr, 2010). The American philosopher and cultural theorist Neil Postman notes that information technologies transform the information society into a society of entertainment, where essence is lost behind form, the human being avoids profound questions about the meaning of life, morality, and ethics, the personality is transformed into an object of consumption, and the connection with history and traditions is lost (Postman, 1985).

A complex, systemic, realistic approach is found in the work of Manuel Castells, who distinguishes such positive aspects as access to knowledge, self-expression and creativity, self-realization through blogs, video, art, flexibility in work, and the possibility of working remotely, while among the threats to human development, he identifies the influence of informational noise and the possibility of manipulation of consciousness (Castells, 1999). An equally objective approach is held by Francis Fukuyama, who analyzes both the influence of information technologies on the development of democracy and the spread of humanistic values, and also points to the potential dangers of technical progress for morality and human nature, and the implementation of biotechnologies (Fukuyama, 1992). Taking the same position, the authors of the article also point to the fact that information technologies positively influence the critical and independently thinking personality, and at the same time, contribute to the further de-intellectualization of the human being of mass culture, the human of the masses (Ozhevan, Dubov, 2017, p. 202). The domestic philosopher V.H. Voronkova identifies as positive aspects the development of education, culture, professional knowledge and experience, and as a negative aspect the intensification of the technogenic character of society (Voronkova, 2012, pp. 21–22). The modern human being, in the pursuit of profit, disregards the preservation of nature, becoming an alternative to Homo ecologicus.

The analytical context of the influence of information technologies is an element of a somewhat different classification, precisely because it not only identifies the very fact and orientation of informational influence, but requires clarification of the reasons for the negative influence of information technologies on the human being. Among the representatives of the analytical approach, it is possible to distinguish the anthropological, informational, socio-political, and technological aspects. Thus, M. Castells, speaking from the standpoint of the anthropological approach, emphasizes that the character of the influence of information technologies depends on the human being

itself – on how he or she is able to use new opportunities for creativity, education, and social interaction, while at the same time avoiding the traps of digital inequality, manipulation, and alienation (Castells, 1999). The Israeli historian and philosopher Yuval Noah Harari, supporting this position, emphasizes that the future of humanity depends on our ability to preserve a balance between the use of technologies and the preservation of humanistic values (Harari, 2017). The domestic philosopher V.H. Kremen also notes that the protection of the human being in the information society lies in self-creation, in the search for meaning and purpose in life, in maintaining dignity, and most importantly, in preserving oneself and one's nature from temptations, which are increasingly numerous in digital reality (Kremen, 2021, p. 20).

The sources of another, informational approach, lie in the works of the French sociologist Jean Baudrillard, who proposed the concept of hyperreality, characterized by the replacement of the real by certain signs: simulacra. Simulations of reality contribute to the inability of people to distinguish reality from fantasy (Baudrillard, 1988). Philosophers also pointed out such important aspects of the influence of information technologies on the human being as "fragmentation of knowledge, when the human being becomes a part of the global flow of information, where changes take place quickly, and stability disappears" (M. Castells, M. Ozhevan, D. Dubov), and the "formation of 'clip consciousness', a fragmented, simplified perception of the world, which reduces the capacity for abstract and consistent thinking (Nicholas Carr, Neil Postman). It is important to draw attention to the opinion of O. Ye. Homilko, who considers the cause of the "weakening of the rational resource" to be precisely the "technologies of producing fakes in the conditions of the modern digital age" (Homilko, 2021, p. 5). An equally interesting direction of research into the causes of personality deformation in the information society lies in the socio-political context, whose representatives connect information technologies with the mode of production, the aspiration for reliable control over the masses, which has direct relevance to the human being. Thus, the British philosopher Mark Fisher notes that the capitalist system, reinforced by digital technologies, creates a "depressive culture" and contributes to social passivity (Fisher, 2010), while the professor of Harvard Business School, social psychologist and philosopher Shoshana Zuboff, expresses the opinion that it is precisely modern monopolies that establish total surveillance over each consumer, controlling human behavior (Zuboff, 2019). M. A. Kholodna sees the cause of such total surveillance in the fact that "no totalitarian regime is interested in developing the intellectual capacities of its citizens, since it is much easier to govern unintelligent people" (Kholodna, 2024, p. 9). Finally, a large number of researchers see the causes of the transformation of the human being as a result of the uncontrolled process of technological singularity. Thus, V.H. Voronkova, emphasizing that "no one will any longer be able to stop this technologization, which is gaining ever greater scale" (Voronkova, 2012, p. 22), concludes that "global informatization is only a component part of the technical and technological transformation of society and permeates the fabric of earthly life with informational networks" (Voronkova, 2012, p. 20). The pace of economic production is increasing to such an extent that it threatens life itself, as well as the spiritual and psychological health of the human being.

Thus, we see quite a broad spectrum of different approaches concerning both the nature of the influence of technologies on the human being and its development, as well as the analysis of causes of both an objective and subjective character, which radically change the nature of human existence. At the same time, it appears expedient to identify not only the nature and causes of the influence of information technologies on the personality, but also their consequences – to comprehend the specificity of the modern socialization of the personality as a type of homo sapiens, and to envision the general tendency that leads to the formation of new images (models) of personality.

3. Results and Discussion

According to the reports of the ICT Data and Statistics Division (ICT – information and communication technologies) of the International Telecommunication Union (ITU), in recent years the world has witnessed the continuous expansion of ICT services and the acceleration of their implementation worldwide. The amount of time that a person spends in information networks is also continuously increasing. Therefore, the general logic of personality transformation is examined in this article by tracing the movement of humanity from the basic model 1.0 to model 3.0 along the time axis, where:

- model 1.0 – partial immersion in the internet space] the vast majority of society has experienced this stage, which society has already passed;
- model 2.0 – complete immersion in the Internet space and deviant changes in the nature of the human being; a small part of society has reached this stage, which is only beginning; in fact, we are now transitioning to a new social condition and the real transformation of the human being;
- model 3.0 – the stage which reflects the relatively near future of the formation of the post-human, involving the synthesis of human and artificial intelligence (see Table 2).

Basic personality models	Subsystems of basic models
Model 1.0 – Homo Informaticus	Model 1.1 Homo Digitalis
	Model 1.2 Homo Irretius
	Model 1.3 Homo Consumens
Model 2.0 – Homo Virtus	Model 2.1 Homo Ludens Digitalis
	Model 2.2 Homo Connectus
	Model 2.3 Homo Algorithmicus
Model 3.0 – Homo Ciberius	Model 3.1 Homo Hybridus
	Model 3.2 Homo Deus
	Model 3.3 Homo Electricus

Table 2. Main models of personality in the information society.

A deeper understanding of the nature of transformational processes will be facilitated by the examination of the subsystem of basic models, which are in fact considered

their structural components. All these additional models have already been proposed by contemporary researchers, but it was necessary to organize them in some way: to determine their place on the time axis, to clarify their meaning, and to show their interconnection with the basic models, and with each other.

The starting point of our research is the awareness of the position of the most widespread stratum of the population, the level of its intellectual development. To do this is not simple, for if civilization, which should in general be considered as a way of the existence of reason in the Universe (Mishchenko, 2024, p. 24), is experiencing the third stage of the intellectual revolution, then it would seem that the contemporary, most widespread human being could be called Homo intelligens - "the knowing human," as was proposed as early as the 1980s by the Japanese theorist of the information society, Yoneji Masuda, who believed that precisely such a human being – Homo intelligens – would ensure revolutionary transformations in the field of information and technologies and embody within itself the creator and bearer of information and knowledge (Masuda, 1985). However, humanity already has a certain experience of existence in contemporary society, which we call informational, and over the past nearly half a century of the formation of information technologies it is possible to draw certain conclusions. Thus, on the basis of our analysis, it is possible to see that information technologies not only provide opportunities for the creative self-expression of the personality, as indicated by representatives of the optimistic approach, including R. Kurzweil, A. Toffler, B. Gates, P. Lévy, and Y. Masuda, but also undermine critical and abstract thinking, as argued by N. Carr, N. Postman, N. Harari, and O. Homilko. The fact that we have not yet created the conditions for the emergence of Homo intelligens also indicates that the transformation of the contemporary information society into a "knowledge society," as envisioned by V.Yu. Bykov (Bykov, 2012), or into a "civilization of intellectual creation," as proposed by V.A. Kipensky and M.V. Smoliaga (Kipensky, Smoliaga, 2023), remains an aspiration.

Thus, under conditions where informatization is considered as "the main tendency of the socio-natural evolution of humanity" (Danyl'yan, Dz'oban, 2019, p. 14), the contemporary human being is the characteristic representative of this actually existing society; it is Homo informaticus. This conclusion is confirmed by statistics: "more than 87% of people, as noted by Glib Vishlinsky, work daily on the Internet" (Vyshlinsky, 2013); they have become accustomed to considering information as a key resource in human life. Such a human being is characterized by a rather high level of erudition, the presence of general, more often superficial, knowledge in various spheres of social life, and a tendency toward informational activity. In this aspect, the human being uses the advantages and opportunities opened by informational resources for the optimization and improvement of professional and personal qualities, and for the possibilities of creative self-development. At the same time, it has certain problems and deviations, which are revealed in its subsystem organizational structures.

Model 1.1 – Homo Digitalis (the Digital Human), is defining for Homo informaticus. The creation of this model involved Don Tapscott, Luciano Floridi, and Björn Schulz. Homo Digitalis is the human being who not only emphasizes involvement in informational flows, but also reflects the special property of the cybernetic process

– the tendency toward digitalization, as the mathematical representation of objective processes of reality, of the internal mechanisms of the functioning of various systems. Mathematics and informatics play an important role in contemporary society. At the same time, it is necessary to take into account certain limitations: given their structure, not all objects can be formalized. In addition, we must take into account the concept of instability of I. Prigogine, who requires consideration of the chaosogenic factors of reality. That is, Homo Digitalis must preserve the essentially human intuitive and imaginative penetration into essence in the process of cognition. There is yet another side of the problem of digitalization – namely, the gradual loss of abstract, profound, critical thinking by Homo Digitalis as a consequence of the specificity of professional informational-digital activity with large volumes and high speeds of information processing, which is characterized by superficial perception and short-term attention. The existence of such processes is evidenced by the work of Nicholas Carr, "Does Google Make Us Stupid?", in which he analyzes how information technologies change human cognitive capacities, promoting superficiality of thought caused by the excess of information (Carr, 2008).

Model 1.2, "Homo Irretius," is the next structural element of the Homo informaticus model, representing its manifestation as a networked personality. M. Castells drew attention to this aspect, indicating that in the information world, the human being becomes a "node" in information networks (Castells, 1999). Regarding existence in the network as a "new mode of human being," V.H. Kremen writes, emphasizing the fact that "today to be means to be in the network" (Kremen, 2021, p. 12), while O. P. Dzioban and S. B. Zhdanenko conclude that network culture contributes to the emergence of a new type of personality – the "network-information" personality – which is integrated into a special sphere of communications (Dzioban, Zhdanenko, 2020). The term "Homo Irretius," literally translated as "the human caught in the network" (from Latin "irretius" – "caught in a net"), was proposed by the Polish futurologist Edwin Bendik. As Shoshana Zuboff emphasizes, such a networked consumer of information becomes an object of observation and analysis (Zuboff, 2019). The behavior of Homo Irretius is subject to control, analysis, and prediction, raising questions about the personal freedom of man and the possibility of interference in their life. Homo Irretius is caught not only in informational networks but also in the network of manipulative technologies of the Gutenberg Galaxy, oriented toward post-truth politics, with "radical distrust of any Reason …, where technologies in their development begin to outpace morality, science, and religion, which are increasingly relativized" (Ozhevan, Dubov, 2017, p. 32).

Homo consumens is the next model – 1.3 in relation to Homo informaticus – which is not only oriented toward the creation and exchange of information, as considered by Manuel Castells, Alvin Toffler, Michel Foucault, and Jacques Derrida, but primarily is a consumer of information. This aspect of the information society was highlighted by Jean Baudrillard, Herbert Simon, and Shoshana Zuboff. In the current era of consumerism, Homo economicus undergoes a certain transformation, giving rise to a social stratum of netocrats – a new dominant class of our time, which becomes the controlling force in the information society and governs the consumptarians, who

consume information without producing any new values and obey the will of the netocrats (Danyl'yan, Dz'oban, 2019).

Thus, considering the deviant changes in the existence of Model 1.0 Homo informaticus, attention should be paid to the deterioration of abstract thinking in Homo Digitalis, the networked mode of existence of Homo Irretius, which threatens personal freedom, leads to a certain internet dependency, reduces openness to the world, diminishes awareness of real-life problems, increases the risk of manipulation of consciousness, and in the social aspect, the formation of a new type of social governance occurs, with the domination of Homo consumens (consumptarians) by the new dominant force of netocrats. In addition to economic dependence, man falls under the power of manipulative technologies.

To observe the further transformation of Model 1.0 Homo informaticus, it is sufficient to analyze these tendencies and focus on the new generation emerging within Model 1.0 – the so-called Generation Z, or the transitional model Homo Zappiens (Wim, Vrakking, 2006). D. Shevchuk notes that the generation tied to television is being replaced by a generation tied to the Internet, the human being ceases to be bound to the physical world, creating an electronic double, a digital avatar, sometimes called the electronic soul of the modern human (Shevchuk, 2011, p. 14). This describes a generation that has grown up under digital technologies and exhibits not only deviations related to the loss of the ability for deep analysis, critical thinking, and concentration (cognitive superficiality), but also a lack of long-term planning and the capacity to work on complex tasks, increasing loneliness, absence of empathy, and a decrease in practical skills for solving real problems due to excessive dependence on digital technologies and the loss of connection with reality as a result of immersion in the virtual world. It is not accidental that A.K. Chaplygin expresses the opinion that the current development of students' intellect is unsatisfactory, and "the creative revolution seems even further from us than …25–30 years ago" (Chaplygin, 2016, p. 26–27).

Information technologies create a new culture of human existence, which D. Shevchuk proposes to call either electronic, because it is precisely electronic means of information that create culture 2.0, which is an alternative to traditional cultural institutions, or to call it virtual culture (Shevchuk, 2011, p. 14). M. Castells, drawing attention to the fact that the culture of the information society is built on communication means controlled by electronics, still defines it as a "culture of real virtuality" (Castells, 1999). The virtual is no longer only informational; it is distinct from the ideal and in some way touches the world of ideas, the spiritual world of humans. Some researchers also distinguish Alpha and Beta generations, in which the qualities of abstract thinking deteriorate significantly and openness to the world is absent, and designate such a type of personality with the index 2.0.

An interesting and in some ways useful approach in this regard is to refer to the etymology of the term "virtus" and analyze its subsequent development. In ancient philosophy, it was used both to denote potential, a certain vector of development, and to reveal the hidden active inner force of all being, and as a category of ethics, which corresponded to the concept of "valor, virtue, excellence." It is in this latter sense that

Nicholas of Cusa used the term, understanding it as the dynamism, the capacity of an idea to act. As a paradoxical world of the existence of objects, virtuality was considered by Leibniz. Gilles Deleuze goes further, understanding the virtual as a "space of simulacra," which he interpreted as a "phantasmagoric image devoid of similarity" (Pal'chyns'ka, 2016, p. 214–220). The term "virtus," and with it the understanding of the virtual, carries within it both a potential for creation and the energy to restructure the world, to create something unexpected, unforeseen, extraordinary.

Thus, it is more important to indicate not by what technical means a new reality may be created, but in what world the human being may exist. The transformation from Model 1.0 Homo informaticus is therefore carried out in the direction of further virtualization, the acquisition of a virtual mode of existence. This is a new quality of the human being, which already differs noticeably from Homo informaticus and represents Model 2.0 – Homo Virtus.

It should, however, be taken into account that the true virtual world is real potentiality and possibility, which can be realized if at the turn of the century a spiritual human being is created, one who strives to transform the world, to spiritualize nature. But the contemporary mass human is a person of mass culture, a person of the masses, as Ortega y Gasset wrote. This person does not strive for creation. The contemporary human is a consumer, and such a human – Homo consumens – contributes to the further transformation of the human being, which turns into Homo Virtus, where the virtual retains only one meaning: the detachment of the human being from the real world, where important events occur and significant contradictions and problems exist.

A certain transformation of Homo Digitalis into Homo Ludens Digitalis occurs. O. H. Danylian and O. P. Dzioban note that "the human of the information society immerses into the virtual reality of simulations and increasingly perceives the world as a play environment – Homo Ludens Digitalis" (Danylian, Dzioban, 2019, p. 12). The model Homo Ludens Digitalis, considered as the first structural component of Homo Virtus (as model 2.1) is based on concepts developed in the context of digital technologies and continues the ideas set forth by Johan Huizinga, further developed by Janet Murray, Steven Conway, and Miguel Sicart. The game factor can also be considered an additional intellectual resource for creative search, but in the context of the pursuit of entertainment, contemporary Homo Ludens Digitalis not only faces a growing risk of "escape" from reality but also the development of deviation into pathology: the formation of gaming addiction as a form of substance dependency.

Similarly, Model 1.2, Homo Irretius, transforms into Model 2.2, Homo Connectus (the Connected Human). If Homo Irretius is a human merely caught in the network, then Homo Connectus describes a human already integrated into global networks and living in a world where connections among humans, technologies, and information become key elements of identity and existence. Luciano Floridi emphasizes that human connections with information systems become fundamental to human existence and notes that it is precisely through these connections that physical and digital realities merge (Floridi, 2014). Constant connection to digital and informational networks induces in Homo Connectus a sense of loneliness, detaches the individual from the

real world and nature, and substantially transforms their spirituality. This is significant because it is precisely a sense of sanctity and awe before the marvelous structure of the universe that forms spirituality, understood as the ability "to translate the universe of external being into the internal universe of the personality on an ethical basis ...leading to meaningful cosmogony, connecting the image of the world with the moral law of the personality" (Kryms'kyy, Pavlenko, 2007, p. 50). Such a movement toward the loss of spirituality significantly diminishes creative qualities, making the human being more pragmatic but less open to the world and to the search for heuristic solutions.

In addition to the deformation of spirituality, Homo Virtus, who constantly works with computers and trusts them because of knowledge of their intellectual capacities, gradually loses independence in making intellectual decisions and intellectual leadership in interaction with artificial intelligence. This transforms into Model 2.3 Homo Algorithmicus (the Algorithmic Human), which emerged in the context of research on the influence of algorithms and big data on the consciousness, behavior, and thinking of the contemporary human. The study of such a human, whose life is increasingly determined by algorithms, artificial intelligence, and automated systems, was conducted by Yuval Noah Harari, while Antoine Pico points out that such a human lives in a world where algorithms play a central role in shaping decisions, behavior, and worldview, and under the influence of algorithmic machine thinking, human thinking, ethics, and culture change (Piko, 2022).

Thus, gaming addiction of Homo Ludens Digitalis, narrowness (closedness to the world) and deformation of spirituality in Homo Connectus, and algorithmic and superficial thinking in Homo Algorithmicus allow Homo Virtus to be considered a transitional model of Homo sapiens, which has already undergone significant deviations and is unable to utilize the intellectual power of the virtual space to transform the real world, because it does not realize the need to spiritualize nature, to fulfill its human mission in the universe, nor even the processes occurring within itself. Such a human gradually transforms into a posthuman, or Model 3.0 Homo Cyberus. An example of contemporary experiments is the search for the creation of an integrated brain-machine interface platform with thousands of channels by Elon Musk (Musk, 2019). Observing the fact that the future of the human-cyborg is already being realized today, O.E. Radutny notes that the creation of living beings Homo Cyberus, which combine organic and inorganic components (so-called "cyborgs" – biological organisms "containing mechanical and/or electronic components"), is no longer science fiction (Radutny, 2018, p. 162).

What is common to all multidimensional modifications of posthumans (cyborgs, clones, mutants) is the understanding of the general contours of this model, or a certain image of personality that unites all these varieties. An example of such a characterization is provided by O. H. Danylian and O. P. Dzioban, who argue that such a human is less burdened by historically conditioned stereotypes, possesses more mobile reactions and the ability to manipulate any layers of information, and acquires a new abstract form of existence: its individuality is lost, and only roles remain (Danylian, Dzioban, 2019, p. 12). One may agree with this definition, but Homo informaticus was

already considered a new quality of personality, whereas Homo Virtus differs by a new mode of being, new sociocultural conditions of existence, a person becomes part of cyberculture (Levy, 1999). Homo Cyberus, however, represents the final transition beyond human abilities and qualities, where we can no longer understand who it really is. "The posthuman," as V.H. Kremen believes, "is a human of information-computer reality, for whom anthropomorphism and humanity are of no significant importance" (Kremen, 2021, p. 22). Therefore, today the question becomes relevant: "Where and when does the 'I' of a human begin and end, and where and when does it begin and end as a personality?" (Kremen, 2021, p. 25).

If Homo Virtus can be considered a deviant, or even pathological (sick) human, Homo Cyberus is regarded not only as a consequence of artificial, voluntary, or forced intervention in its organic nature, but as a being that will not necessarily retain even the innovative spirituality that still existed in Homo Virtus in its striving for self-development (it is not accidental that A. Maslow considers the pursuit of self-realization as the most important for humans of this era, as a manifestation of their own spiritual development). In the case of Homo Cyberus, the situation may change. The human abandons the pursuit of intellectual self-improvement and seeks to expand its intellectual capabilities through bioengineering and cybernetics, forming Model 3.1 Homo Hybridus, a being aiming to enhance human cognitive functions, the ability to process information faster, and make decisions through the implantation of certain devices into the human brain and the formation of specific neurointerfaces, technologies that create certain augmented reality (AR), enriching human perception of the real world with digital data, but still preserving the drive for self-development as a source of spirituality, while destroying human individuality. The human begins to define itself not only as a biological entity but also through a digital profile and integration with technologies. According to Donna Haraway, such a change in self-identity should not only mark an endpoint in human history but also the end of divisions between nature and culture, human and animal, male and female, calling for a rethinking of these categories as outdated in the contemporary world (Haraway, 1991).

What is the further fate of Homo Hybridus, and with it, Homo Ciberus? It is necessary to consider that the process of creating and establishing Homo Ciberus is not simultaneous, but gradual, and humanity must go through the period when, according to M. Beylin, a person "must preserve their human essence, their human substance, their human image in the boundlessness of technotronic civilization with robots, with intellectual intelligence, with the cloning of beings, with the immense horizons of technical possibilities and prospects" (Beylin, 2014, p. 444). In this new society, a ruling class of netocrats is already emerging, and two different alternatives will form the two poles of two diametrically opposed models of Homo Ciberus: 3.2 Homo Deus and 3.3 Homo Electricus.

According to Yuval Noah Harari, Homo Deus will be able to go beyond its biological limitations and acquire qualities previously attributed to gods: immortality, unlimited knowledge, and control over natural processes (Harari, 2017). Homo Deus is not only the fruit of bioengineering and cyborgization, but also the result of cooperation with artificial intelligence, which leads to the convergence of humans and artificial

intelligence. Thus, J. C. R. Licklider in the article "Man-Computer Symbiosis" writes about "coexistence in close connection or even close union of two heterogeneous organisms" (Licklider, 1960). Contemporary Japanese researcher Jun Rekimoto in the work "Homo Cyberneticus: The Era of Human and AI Integration" points out that "instead of considering artificial intelligence as an autonomous or separate entity," it should be regarded as "an integral part of humans" (Rekimoto, 2019).

However, the divine immortality of Homo Deus, as well as its happy future, raises doubts. Happiness itself becomes a scientific problem that can be measured, improved, and controlled. The basis of its worldview becomes technogumanism, and the new form of spirituality – dataism, as "worship of data" and "universal faith in the omnipotence of algorithms" (Harari, 2017). It is also necessary to consider the factor of technological singularity, when automated production will be controlled by artificial intelligence, and humans will lose their leading role and become part of artificial intelligence itself, not only resulting in a multiple increase in the pace of technological development but also in the emergence of an unpredictable vector of this development, which is subject to the technocratic slogan: the development of technology for the sake of the development of technology itself without considering the preservation of life and protection of nature. Humans, as spiritual beings, will effectively lose the meaning of their existence.

The fate of the consumer class, which forms the basis of Homo Electricus (model 3.3), seems equally sad, as it undergoes no less radical transformations. Firstly, this is a form of genetic mutation capable of endlessly developing its own qualities necessary for performing human productive functions and certain social roles (for example, an ideal soldier). Secondly, stimulation of vital activity must occur according to the same principle of seeking pleasure as in Homo Deus, who is focused on optimizing internal states using pharmacology and digital interfaces. Even now, medicine uses deep brain stimulation to treat such diseases as Parkinson's, studying brain areas responsible for feelings of pleasure. In the future, stimulation of work and performance of social roles by Homo Electricus may occur through the illusion of possible pleasure with the release of pleasure hormones (endorphins). Not accidentally, A.V. Kipensky, noting that "artificial brain stimulation to obtain ephemeral sensations is no different from ordinary drug addiction and is extremely irrational from the standpoint of human survival," writes that "one must stop at the last limit, define new humanistic principles and ideals, new spirituality for the further existence and development of human society" (Kipens'kyy, 2022, p. 7). The very fate of humans depends on how deeply the scenarios of their transformation in the digital society are realized and how the strategies for harmonious human development in the 21st century are determined.

4. Conclusion

The development of the information society is a rather contradictory process in human history, posing a threat of an anthropological crisis. Its essence lies in the gradual transformation of humans from Homo informaticus to Homo Virtus and Homo

Ciberus and threatens the end of human evolution as the species Homo sapiens. Revealing the tendencies of such processes to humans themselves may have a significant influence by awakening their self-awareness, but halting such processes through educational work is impossible. The preservation of humanity in humans is possible only if people themselves become agents of creative, actively transformative activity.

Therefore, humanity needs a new development strategy. The transformation of humans should begin with the transformation of the information society itself into a technotronic one, where technology, subordinated to the ecological, demographic, and energy interests of social development, is the determinant of the transformation of the infosphere from a means of entertainment and communication into an intellectual one, based on the knowledge society.

This vision of human development can be summarized in the words of Mishchenko (2023): "The fulfillment of this task activates a person's vital forces, enabling them to contribute to the realization of their own cosmic mission – the spiritualization of the Universe, the terraforming of planetary landscapes, and the maintenance of stable developmental trends. The 'return' of the human being to the real world will undoubtedly be facilitated by the incorporation, within the educational environment, of the idea of spiritual upbringing, whose purpose is to shape the orientation of the person of the future toward discovery and creation, toward a full-fledged existence in the world as an active, creative being." (Mishchenko, 2023).

References

Baudrillard, J. (1988). Simulacra and simulations. In *Selected writings* (pp. 166–184). Stanford University Press.

Beylin, M. V. (2014). *Nanotekhnolohiya yak proryv u postneklasychnoyi nautsi* [Nanotechnology as a breakthrough in post-nonclassical science]. Oberig.

Bykov, V. Yu. (2012). Innovatsiynyy rozvytok zasobiv i tekhnolohiy system vidkrytoyi osvity [Innovative development of tools and technologies of open education systems]. *Suchasni informatsiyni tekhnolohiyi ta innovatsiyni metodyky navchannya u pidhotovtsi fakhivtsiv: Metodolohiya, teoriya, dosvid, problemy, 29*, 32–40.

Carr, N. (2010). *The shallows: What the Internet is doing to our brains*. W. W. Norton & Company.

Castells, M. (1999). *The information age: Economy, society and culture*. Wiley-Blackwell.

Chaplyhyn, A. K. (2016). Nauchnoe y tekhnyko-tekhnolohycheskoe tvorchestvo y budushchee chelovechestva [Scientific and technological creativity and the future of humanity]. In *Teoretychni ta prykladni problemy vzayemodiyi nauky, tekhniky i tekhnolohiyi: Materialy II Vseukrayins'koho naukovo-praktychnoho seminaru, prysvyachenoho 85-richchyu Kharkivs'koho natsional'noho avtomobil'no-dorozhn'oho universytetu* (pp. 25–43). KHNADU; Vydavnytstvo "Lider."

Danyl'yan, O. H., & Dz'oban', O. P. (2019). Lyudyna v informatsiynomu suspil'stvi: Problema moral'noyi identyfikatsiyi [Human in the information society: The problem of moral identification]. *Visnyk Natsional'noho Yurydychnoho Universytetu imeni Yaroslava Mudroho, 1*(40), 8–20.

Dz'oban', O. P., & Zhdanenko, S. B. (2021). Informatsiyna revolyutsiya: Sotsioantropolohichni ta svitohlyadni transformatsiyi [The information revolution: Socio-anthropological and worldview transformations]. *Informatsiya i Pravo, 4*(39), 22–34.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford University Press.

Fisher, M. (2010). *Capitalist realism: Is there no alternative?* Zero Books.

Fukuyama, F. (1992). *The end of history and the last man*. Free Press.

Gates, B. (1999). *Business @ the speed of thought: Succeeding in the digital economy*. Penguin Books.

Harari, Y. N. (2017). *Homo Deus: A brief history of tomorrow*. Harper.

Haraway, D. (1991). A cyborg manifesto: Science, technology, and socialist feminism in the late twentieth century. In *Simians, cyborgs, and women: The reinvention of nature* (pp. 149–181). Routledge.

Homilko, O. Ye. (2021). Filosofiya yak propahanda u tsyfrovykh vyklykakh mrakobissya [Philosophy as propaganda in the digital challenges of obscurantism]. In *Filosofiya v suchasnomu sviti: Materialy II Mizhnarodnoyi naukovo-praktychnoyi konferentsiyi* (pp. 5–8). National Technical University "Kharkiv Polytechnic Institute."

Howe, N., & Strauss, W. (1991). *Generations: The history of America's future, 1584–2069*. William Morrow.

Jaspers, K. (1949). *Vom Ursprung und Ziel der Geschichte*. Fischer Bücherei.

Kholodna, M. A. (n.d.). *Psykholohiya intelektu: Paradoksy doslidzhennya* [Psychology of intelligence: Research paradoxes]. Retrieved October 6, 2025, from file:///D:/1.%20%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82/Holodnaya_Psihologiya_intellekta.pdf

Kipens'kyy, A. V. (2022). Transhumanistychne maybutnye informatsiynoho suspil'stva: Zberezhennya lyuds'koyi pryrody ta dukhovnosti [The transhumanist future of the information society: Preserving human nature and spirituality]. *Visnyk Natsional'noho Tekhnichnoho Universytetu "KhPI". Seriya: Aktual'ni Problemy Rozvytku Ukrayins'koho Suspil'stva, 2*, 5–11.

Kipens'kyy, A. V., & Smolyaha, M. V. (2023). Lohika rozvytku tsyvilizatsiyi: Vid informatsiynoho do intelektual'noho suspil'stva [The logic of civilization's development: From information to intelligent society]. *Doslidzhennya z Istoriyi i Filosofiyi Nauky i Tekhniky, 32*(1), 42–55.

Kremen', V. H. (2021). Lyudyna v informatsiynomu sviti [Human in the information world]. In *Stratehiyi formuvannya tvorchoyi osobystosti: Metody, pryyomy, formy* (pp. 9–29). Instytut obdarovanoyi dytyny NAPN Ukrayiny.

Kryms'kyy, S. B., & Pavlenko, Yu. V. (2007). *Tsyvilizatsiynyy rozvytok lyudstva* [Civilizational development of humanity]. Feniks.

Kurzweil, R., & Grossman, T. (2009). *Transcend: Nine steps to living well forever*. Rodale.

Lévy, P. (1994). *L'intelligence collective: Pour une anthropologie du cyberspace*. La Découverte.

Lévy, P. (1999). *Cibercultura*. Editora 34.

Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics, HFE-1*(1), 4–11.

Masuda, Y. (1985). Hypothesis on the genesis of *Homo intelligens*. *Futures, 17*(5), 492–494.

Miles, S. (1998). *Consumerism as a way of life*. SAGE.

Mishchenko, V. I. (2023). *Intellectual and moral spirituality in a technological society* (Doctoral dissertation). National Technical University "Kharkiv Polytechnic Institute."

Musk, E., et al. (2019). An integrated brain-machine interface platform with thousands of channels. *Journal of Medical Internet Research, 21*(10). https://preprints.jmir.org/preprint/16194

Ozhevan, M., & Dubov, D. (2017). *Homo ex machina: Filosofs'ki, kul'turolohichni ta politychni peredumovy formuvannya konverhentnoho suspil'stva* [Homo ex machina: Philosophical, cultural, and political prerequisites for the formation of a convergent society]. NISD.

Pal'chyns'ka, M. V. (2016). *Virtual'nyy prostir v umovakh sotsiokul'turnykh transformatsiy* [Virtual space under conditions of sociocultural transformations] (Doctoral dissertation, South Ukrainian National Pedagogical University named after K. D. Ushynsky, Odesa).

Piko, A. (2022). *The algorithmic society: Technology, power, and knowledge*. In M. Schuilenburg & R. Peeters (Eds.), Routledge.

Postman, N. (1985). *Amusing ourselves to death: Public discourse in the age of show business*. Viking Penguin.

Radutnyy, O. E. (2018). Tsyfrova lyudyna z tochky zoru zahal'noyi ta informatsiynoyi bezpeky: Filosofs'kyy ta kryminal'no-pravovyy aspekt [The digital human from the perspective of general and information security: Philosophical and criminal law aspects]. *Informatsiya i Pravo, 2*(25), 158–170.

Rekimoto, J. (2019). *Homo cyberneticus: The era of human–AI integration*. arXiv. https://doi.org/10.48550/arXiv.1911.02637

Shevchuk, D. (2024). Kul'tura u merezhyvi merezh [Culture in the network of networks]. *Krytyka, 3–4*, 161–162.

Toffler, A. (1980). *The third wave*. William Morrow.

Veen, W., & Vrakking, B. (2006). *Homo zappiens: Growing up in a digital age*. Continuum.

Voronkova, V. H. (2012). Hlobal'na informatyzatsiya yak holovna tendentsiya sotsiopryrodnoyi evolyutsiyi lyudstva [Global informatization as the main trend of the socio-natural evolution of humanity]. *Humanitarnyy Visnyk Zaporiz'koyi Derzhavnoyi Inzhenernoyi Akademiyi, 51*, 14–26. http://nbuv.gov.ua/UJRN/znpgvzdia_2012_51_4

Vyshlinsky, G. (2013). Ponad 87% ukrayintsiv zakhodyat' v Internet khocha b raz na den' [More than 87% of Ukrainians use the Internet at least once a day]. *MediaSapiens*. http://osvita.mediasapiens.ua/material/28049

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 254 – 270

THE ROOTS OF PRACTICAL REALISM AND METAPHYSICS

Zhumagul Bekenova*

Peeter Müürsepp**

Abstract: The article analyses the roots of the novel understanding of science that has been proposed by the Estonian philosopher of science and chemistry Rein Vihalemm and called practical realism by him. The analysis reveals an interesting intellectual journey that brought a philosopher who was educated in the Soviet Union in alignment with Marxism to quite different ideas of Martin Heidegger, a very popular philosopher in Eastern and Central Europe in the late 20^{th} century. Strangely, studying Heidegger's understanding of modern science helped Vihalemm to re-discover Marx from a new angle relevant from the perspective of the approach to science as a practical activity. The issue of the Kantian influences on practical realism is another focus of the article. Most importantly, the analysis connects to the attitude towards the role of metaphysics in science. Vihalemm accepts Kantianism in its sublated form calling metaphysics exhausted at the same time. The article aims to show that metaphysics still plays an important role in making sense of science, in part as an inherent component of science, and that it is premature to consider metaphysics exhausted.

Keywords: Joseph Rouse, Kantianism, Martin Heidegger, Marxism, metaphysics, practice, realism, Rein Vihalemm.

1. Introduction

Practical realism is a new kind of realist approach in the philosophy of science that was initiated by the late Estonian philosopher of science and chemistry Rein Vihalemm in the first decade of the 21^{st} century. The current analysis aims at taking a deeper look into the main influences Vihalemm encountered along his road to becoming an original thinker in philosophy of science and one of the founders of the new philosophical discipline, philosophy of chemistry.

Our main focus will be on two closely connected topics that we consider important in the context of practical realism: the role of metaphysics and the issue of the Kantian roots of practical realism. Rein Vihalemm himself denied the former and accepted the latter in the 'sublated' form only. Below, we shall explain what the Hegelian term 'sublation' (*aufheben*) actually means and why Vihalemm was eager to apply such terms, which are uncommon in philosophy of science. This is largely due to his preparation as a philosopher in the Soviet context of the field where Hegelian dialectics played a crucial role as the decisive inspiration for Karl Marx in shaping his dialectical materialist view of the world.

*Al-Farabi Kazakh National University, Almaty, Kazakhstan; Bekenova.Zhumagul@kaznu.kz **Tallinn University of Technology, Tallinn, Estonia; peeter.muursepp@taltech.ee

©2026 Bekenova and Müürsepp. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

In Hegel's philosophy, the term *Aufhebung* has two seemingly contradictory meanings of both preserving and changing, followed by an advancement. The idea is very common in German idealism as well as Marxism. In a way, sublation is the engine that drives the dialectical understanding of processes. The approach is based on the struggle (of the opposites) followed by a qualitative leap up to a new level. Below we shall see how Kant's understanding of science (proper) is sublated in practical realism.

Having taken his first academic degree in chemistry, Rein Vihalemm remained influenced by his basic discipline throughout his career as a philosopher and historian of science. This influence can be felt in the core ideas of practical realism as well.

Rein Vihalemm saw practical realism as an alternative to both antirealism and standard scientific realism (Vihalemm, 2015, p. 100). Vihalemm presented the core of practical realism in the form of five tenets:

(1) Science does not represent the world "as it really is" from a god's-eye point of view. Naïve realism and metaphysical realism have assumed the god's-eye point of view, or the possibility of one-to-one representation of reality, as an ideal to be pursued in scientific theories, or even as a true picture in the sciences.

(2) The fact that the world is not accessible independently of scientific theories – or, to be more precise, paradigms (practices) – does not mean that Putnam's internal realism or "radical" social constructivism is acceptable.

(3) Theoretical activity is only one aspect of science; scientific research is a practical activity and its main form is the scientific experiment that takes place in the real world, being a purposeful and critical theory-guided constructive, as well as manipulative, material interference with nature.

(4) Science as practice is also a social-historical activity which means, amongst other things, that scientific practice includes a normative aspect, too. That means, in turn, that the world, as it is accessible to science, is not free from norms either.

(5) Though neither naïve nor metaphysical, it is certainly realism, as it claims that what is "given" in the form of scientific practice is an aspect of the real world. Or, perhaps more precisely, science as practice is a way in which we are engaged with the world (Lõhkivi, Vihalemm, 2012, p. 3).

There is probably no need to explain what antirealism means but we have to give Rein Vihalemm's understanding of standard scientific realism as he called it. Vihalemm put it like this:

By standard scientific realism I mean the conception according to which:

(1) there is a mind-independent world (reality) of observable and unobservable objects (the metaphysical-ontological aspect),

(2) the central notion is truth as correspondence between scientific statements (theories) and reality (the semantic aspect),

(3) it is possible to obtain knowledge about the mind-independent reality (the epistemological aspect),

(4) truth is an essential aim of scientific inquiry (the methodological aspect) (Vihalemm, 2015, p. 100).

According to Vihalemm, being an antirealist means rejecting at least one of these points. Therefore, a practical realist cannot reject any of these four points but can reinterpret them. Vihalemm claimed that practical realism avoids the shortcomings of both anti-realism and standard scientific realism. Considering Vihalemm's understanding of the latter, it is clear that the successful avoidance of these shortcomings must have a lot to do with metaphysics and truth. Truth is deflationary for him as we'll see shortly. The issue of metaphysics, however, is more complicated. To begin with, we claim that the metaphysical-ontological aspect has to be interpreted in a way that enables us to avoid the God's-eye view syndrome. Making metaphysics deflationary does not work for Vihalemm. He tries to do away with metaphysics altogether. However, accepting the deflationary approach to truth is hardly enough for fulfilling this job. More is needed, but can this be achieved? We'll analyse below. There seems to be no special problem with the epistemological aspect of standard scientific realism unless we accept a type of realism that assumes the God's-eye view.

The focus of the issue of avoiding the shortcomings of both anti-realism and standard scientific realism is on the role of truth and metaphysics, as pointed out above. In the current analysis, we shall concentrate mostly on metaphysics. Concerning the attitude towards truth, we acknowledge the position of Rein Vihalemm who declared that he is ready to assume the deflationary attitude to truth: "I cannot speak for pragmatists, but in practical realism, 'truth' can be interpreted in a deflationary way and this interpretation is compatible with semantic realism" (Vihalemm, 2011, p. 55). We reserve further analysis in this direction for forthcoming papers. However, we must also make some remarks here, mostly because the deflationary option becomes important in the case of metaphysics as well. It is important to add, although this is not directly the focus of this paper, that standard scientific realism suffers from the lack of emphasising the practical aspect of scientific research.

As emphasised above, we concentrate on the issue of metaphysics in practical realism and the question of what Kantianism has to do with this. Rein Vihalemm ended one of his publications with the statement: "Metaphysics is exhausted but practical realist thinking remains, to be appreciated and furthered" (Vihalemm, 2013, p. 12). In the same paper, Vihalemm is analysing the impact of Kantianism in practical realism building a connection between the issue of metaphysics in science and Kant's conception of proper science. This is the reason for us to analyse these problems in conjunction as well. After all, it is part of the most basic knowledge of philosophy that Kant put forward his idea of proper science in the context of attempting to make metaphysics real science. Vihalemm's aim was opposite in a way, explaining the idea of proper science without any connection to metaphysics whatsoever but still strongly acknowledging the role of mathematics. The latter seems to be the core of Vihalemm's understanding of sublating Kant. The basic approach, the need for the mathematical

basis is kept but, at the same time, the need for keeping metaphysics is neglected altogether. However, we are still left with the question: why is metaphysics exhausted? How exactly should we understand this claim? The answer to this question requires taking a detailed look into the development of Rein Vihalemm as a philosopher of science.

As a kind of warning to the reader, we are going to address topics and thinkers that are not often at stake for philosophers of science. However, there is no other way to properly understand the background and essence of Rein Vihalemm's conception of φ-science (a model of science proper based on Galilean physics) and practical realism. Vihalemm matured as a philosopher within the Marxist-dominated Soviet system and was able to engage with Western philosophical traditions only in the middle of his career. Consequently, he has had to study philosophers, like Hegel, whose works are normally quite unknown to Western philosophers of science. Vihalemm's interest in Kant and Heidegger, not to speak about Marx, is also strongly related to his background and the social-historical circumstances he had to live through during his active years as a philosopher and historian of science. All thinkers just mentioned have made their mark on the thought of Rein Vihalemm. Our claim is, however, that the influence of Immanuel Kant laid the decisive foundation for the understanding of the essence of science for Vihalemm.

The special puzzle for us derives from the fact that Kant was aiming to make metaphysics into a science. Rein Vihalemm, however, did not recognise any role for metaphysics in connection with science proper in any sense, but still considered a clear answer to the main question of philosophy of science – what is this thing called science – vitally important. Obviously, this answer can be given only on the basis of a justified explanation of what kind of science we should understand as science proper. Is it just physics, or perhaps chemistry, or do we need an explanatory model? Vihalemm's response is the latter, the model of φ-science, that is still strongly based on physics as even the name of the concept suggests. The conception of φ-science has been recently thoroughly analysed from different angles by Ave Mets (2024a, pp. 6–36; 2024b, pp. 3–31; 2025, pp. 3–30).

2. From Marx to Heidegger and back again

Rein Vihalemm matured as a philosopher in the Soviet Union where Marxism was officially acknowledged as the only correct view of the world. It can be argued that the official ideological doctrine in the Soviet Union was not really a version of original Marxism, but this would be a different story, although perhaps quite an interesting one. Everyone involved in academic philosophy had to at least pretend to be Marxist. In informal communication with each other, Estonian philosophers of that period used the concept of 'foreword Marxism' as it was common to cite the classics of Marxism-Leninism in only the foreword of a philosophical treatise. Normally, the censors did not bother to read any further.

The philosophers of science managed to turn the situation in their favour by concentrating on Karl Marx's approach to practice and applying it to the analysis of

the methodology of scientific research (see Mets, 2019). The main source was Marx's short philosophical piece called, "Theses on Feuerbach". Normally, the first thesis has been pointed out as the core of the issue, thoughts leading to the practical approach to science: "The main defect of all hitherto-existing materialism – that of Feuerbach included – is that the Object [*der Gegenstand*], actuality, sensuousness, are conceived only in the form of the object [*Objekts*], or of contemplation [*Anschauung*], but not as human sensuous activity, practice [*Praxis*], not subjectively. Hence it happened that the active side, in opposition to materialism, as developed by idealism – but only abstractly, since, of course, idealism does not know real, sensuous activity as such. Feuerbach wants sensuous objects [*Objekte*], differentiated from thought-objects, but he does not conceive human activity itself as objective [*gegenständliche*] activity. In The Essence of Christianity [*Das Wesen des Christenthums*], he therefore regards the theoretical attitude as the only genuinely human attitude, while practice is conceived and defined only in its dirty-Jewish form of appearance [*Erscheinungsform*]. Hence he does not grasp the significance of 'revolutionary', of 'practical-critical', activity" (Marx, 1845).

Karl Marx does not speak about science in particular here. His approach to practice is a general epistemological one. However, we don't have any other approach to practice in the context of doing scientific research either. Marx was interested in the interaction between objective and subjective reality. It is not about practice being the criterion of truth that the simplified interpretation suggests. Marx is emphasizing the focal role of practice in making human cognition work and yield adequate information about the world out there. Besides every other walk of life, science definitely does this.

In the second thesis of Marx on Feuerbach the idea of practice as the criterion of truth still comes in: "The question whether objective truth can be attributed to human thinking is not a question of theory but is a practical question. Man must prove the truth, *i.e.*, the reality and power, the this-sidedness [*Diesseitigkeit*] of his thinking, in practice. The dispute over the reality or non-reality of thinking which is isolated from practice is a purely scholastic question" (Marx, 1845).

Reading the second thesis in the context of our analysis leaves us with the question – is it here that Rein Vihalemm's naturalist approach to science originates from? Is it here where the exhaustion of metaphysics started from his perspective? There is no direct answer to be found, but this seems quite possible. However, Vihalemm did not adhere just to the Marxist position in the connection of the issues addressed in the second thesis. The attitude towards objective truth makes a difference. For Marx, achieving objective truth remains the main goal of human cognition. In practical realism, however, the issue of truth deserves almost no attention whatsoever.

Knowledge of Marxism did not guide Rein Vihalemm directly to Kant, at least in the context of practical realism. Somewhat strangely, Martin Heidegger was needed to serve as the mediator for Vihalemm for returning to both Marx and Kant from a fresh viewpoint. For obvious political reasons, Marxism was not fashionable in the former Soviet space in 1990s when Rein Vihalemm started to develop the ideas that eventually led him to practical realism. On the other hand, the philosophy of Martin Heidegger became very popular. Many philosophers in the territory of the former

Soviet bloc were probably looking for an alternative to Marxism. As most of them had been trained in the Hegelian-Marxist rather than in the analytic tradition, Heidegger was a logical choice. Heidegger's philosophy seemed to offer a breath of fresh air that was accessible to thinkers trained in the earlier German philosophical tradition. Nevertheless, Heidegger has since fallen out of fashion, perhaps largely because of the repercussions surrounding the publication of the *Black Notebooks*. In fact, however, interest in Heidegger had already begun to decline before their publication.

Let us note that Rein Vihalemm, not having been fluent in German, interpreted Heidegger with the help of some Anglophone thinkers rather than addressing the German existentialist directly. From the original texts of Heidegger, Vihalemm most probably was mostly influenced by the "Age of the World Picture" where Heidegger positions modern science within his general outlook at the basic areas of human life in the contemporary world. The following observation is guesswork as no direct references in Vihalemm's writings are to be found. However, it may well be that Heidegger's explanation of the role of mathematics in contemporary science (physics) made its decisive mark: "Modern physics is called mathematical because, in a remarkable way, it makes use of quite specific mathematics. But it can proceed mathematically in this way only because, in a deeper sense, it is already itself mathematical" (Heidegger, 1977, p. 118).

In what sense is modern science essentially mathematical? The answer can be found in the meaning of the original Greek term for mathematics: "*Ta mathēmata* means for the Greeks that which man knows in advance in his observation of whatever is and in his intercourse with things: the corporeality of bodies, the vegetable character of plants, the animality of animals, the humanness of man" (Heidegger, 1977, p. 118). Mathematics was the basis of understanding of all things for the Greeks already. In modern science, mathematics has kept its position, but in a much more narrow and concrete sense, as just the language of science. This change is part of the broader shift from qualitative to quantitative thinking about nature that happened in early Modernity. We can see two important points here for the development of the thought of Rein Vihalemm. First, just like in Kant, there is the fundamental position of mathematics in Heidegger's understanding of modern science. Second, the mathematical nature of our understanding of reality is, in a sense, pre-given. For both Heidegger and Kant, human scientific cognition is fundamentally mathematical. However, we cannot say that it is *a priori* in the Kantian sense for Heidegger.

Rein Vihalemm definitely read the "Question Concerning Technology" as well. This is one of the first Heidegger pieces that was translated into Estonian (Heidegger, 1989) and influenced many local philosophers despite the somewhat peculiar usage of language by the translator. (An alternative option was to consult the Russian translation for guidance.)

The main sources on Heidegger in the Anglophone literature for Vihalemm were the works of Jeff Kochan and Trish Glazebrook. With the help of Kochan, Vihalemm claimed that Heidegger's *Dasein* [Being-in-the-World] can be interpreted as practice. Kochan explains: "On Heidegger's account [...] the root problem of both realism and antirealism is that they both begin their analysis *one level too late*. They start

at the level of subject and object rather than at the more primitive level of Dasein [Being-in-the-World] and world. Insofar as both realists and antirealists rely upon a theoretical conception of the world as reality, they both fail to get at the real in a way which discloses entities in their more fundamental, nonobjectival state" (Kochan, 2011, p. 89).

It may seem insufficiently justified to characterize Heidegger's *Dasein* in terms of practice. However, it should be noted that Kochan interprets Heidegger with Joseph Rouse's ideas in mind. Rouse is a leading figure in the practice-based approach to science and also strongly influenced Rein Vihalemm, particularly as a thinker who assigns no role to metaphysics in understanding reality. This is how Rouse sees the core of the issue: "[t]he question is not how we get from a linguistic representation of the world to the world represented. We are already engaged with the world in practical activity, and the world simply is what we are involved with. The question of access to the world, to which the appeal to observation was a response, never arises. The important categories for characterizing the ways the world becomes manifest to us are therefore not the observable and unobservable. We must ask instead about what is available to be used, what we have to take account of in using it, and what we are aiming toward as a goal" (Rouse, 1987, p. 143).

We can see that there is a similarity between Heidegger's *Dasein* and Rouse's being engaged with the world in practical activity. However, Heidegger himself would hardly explain his basic category in such a straightforward way. Therefore, we should not claim that these concepts share exactly the same meaning.

Trish Glazebrook's contribution to Vihalemm's approach lies in establishing a connection between Heidegger's views and realism: "Heidegger is a realist about beings but an anti-realist about being" (Glazebrook, 2001, p. 362). According to Glazebrook, Heidegger can be regarded as a robust scientific realist. If, as Kochan argues, practice plays an important role in Heidegger's thought, a meaningful connection can be drawn between Heidegger's philosophy and practical realism. Rein Vihalemm is likewise sympathetic to interpreting Heidegger as a realist, particularly on the basis of Heidegger's ontological distinction between beings and Being.

Heidegger's importance to philosophy of science has often been overlooked. In this context, it is worth noting Heidegger's remarks, in one of his last papers, on philosophy's transformation into the sciences. He explains that the end of philosophy entails its development into the sciences, which brings the history of philosophy to its fullest possibility, while leaving a task that remains for thinking (Vihalemm, 2013, p. 11). In Heidegger's own words there is a task "accessible neither to philosophy as metaphysics nor, even less, to the sciences stemming from philosophy" (Heidegger, 2011, p. 314).

There is a clear and forceful critique of metaphysics here, adding further support to the observation that philosophers themselves have often been among its fiercest critics. Rein Vihalemm, given the naturalistic orientation of his approach to science, is clearly sympathetic to this aspect of Heidegger's thought. There is more than that. Heidegger claims: "[W]hat we say about the end of philosophy means the completion

of metaphysics" (Heidegger, 2011, p. 312). [...] "With the reversal of metaphysics that was already accomplished by Karl Marx, the uttermost possibility of philosophy is attained" (Heidegger, 2011, p. 313). As we can see, Heidegger gives full credit to Karl Marx as the thinker who initiated a new approach to metaphysics.

There are two important points here that favour the practical realist approach. Metaphysics is about to end according to Heidegger, and this development is due to philosophy evolving into the sciences. According to Vihalemm, metaphysics has to be kept out of science. Vihalemm's hope was that this will exactly be one of the most important tasks of practical realism, and there are all reasons to believe in the successful outcome.

In addition, as we just observed, Heidegger seems to be thinking that Marx liberated philosophy from metaphysics. The claim may really hold if we look at the second thesis of Marx on Feuerbach.

Ultimately, we arrive at the somewhat unexpected conclusion that, in a sense, Martin Heidegger led Rein Vihalemm back toward Marxism, with some assistance from Jeff Kochan and Trish Glazebrook. But what about Kant? How did he reappear in Vihalemm's thinking? Rein Vihalemm explains: "Heidegger agrees with Kant in that whatever one must presuppose in order to know any object whatsoever, cannot itself be known as an object (*Objekt*; A 402), but disagrees with Kant that these presuppositions must be transcendentally ideal; instead, Heidegger contends, they are ways of being in the world, that is, ways we behave, act and find ourselves within our surroundings" (2013, p. 11).[1]

We have no reason to claim that Heidegger's connection to Kant becomes clear on the basis of this quote. If we try to explain in greater detail: Heidegger agrees with Kant in that whatever one must presuppose in order to know any object whatsoever cannot itself be known as an object. Heidegger disagrees with Kant that these presuppositions must be transcendentally ideal. Heidegger contends, they are ways of being in the world, that is, ways we behave, act and find ourselves within our surroundings (practices). The same disagreement is there in practical realism. Vihalemm needed transcendental idealism to be sublated (*aufheben*). Below, we shall explain the meaning of this Hegelian concept and its application by Vihalemm in more detail.

It is obvious that studying Heidegger at the time of the peak of the popularity of the German thinker in the former Soviet bloc enabled Rein Vihalemm to discover important moments from the point of view of philosophy of science in the thought of Immanuel Kant and Karl Marx. It may well be that without these connections, practical realism would never have been born. It appears that practical social circumstances have played their remote role in the formation of a philosophical approach to science.

Let us now take an even deeper look into the suggested Kantian roots of practical realism.

[1] Rein Vihalemm has acknowledged the contribution of Kenneth Westphal to his understanding of Heidegger's interpretation of Kant.

3. Immanuel Kant's conception of proper science and practical realism

Rein Vihalemm's practical realism represents a further development of his own conception of proper science, which he called *φ*-science. We claim that the latter is strongly influenced by Kant's conception of proper science. In a way, it can even be called a contemporary specification of Kant's approach to the matter. Still, there is an important philosophical difference between the two as far as the aims are concerned. Vihalemm did not aim at an idealist but rather a practical view of the matter.

What did Kant mean by proper science (*eigentliche Wissenschaft*)? Kant's quite well-known explanation reads: "I assert [...] that in any special doctrine of nature there can be only as much *proper* science [*eigentliche Wissenschaft*] as there is *mathematics* therein. For, according to the foregoing, proper science, and above all proper natural science, requires a pure part lying at the basis of the empirical part, and resting on a priori cognition of natural things" (Kant, 2002b[1786], p. 4, p. 470). Let us remember that according to Heidegger, modern physics is mathematical by its essence.

Kant suggested that mathematics provides us with *a priori* principles for cognizing physical laws. Mathematics is essential for Rein Vihalemm's model of *φ*-science as well. Still, Vihalemm claims that Kant's transcendental idealism and apriorism as premises of proper scientific knowledge can be overcome within practical realism (2013, p. 7). Here, Vihalemm directly applies the Hegelian idea of 'sublation' (*Aufhebung*). The term has a threefold meaning: "to cancel", "to preserve" and "to elevate", something very typical for Hegel's triadic dialectical thinking.

Let us note that knowledge of Hegelian dialectics was a required part of the education of philosophers in the Soviet system for an obvious reason. Hegel's dialectics was considered vitally important as the basis of Karl Marx's dialectical materialism, his theoretical worldview. The three laws of dialectics – the struggle of the opposites, change of quantity into quality (qualitative leap) and negation of negation – were picked up by the Marxists as the general explanation of not only natural but of social development as well. It is important to notice that there are exactly three basic laws of the dialectics. For Hegel, the triad – thesis, antithesis and synthesis – was the framework for understanding reality in most general terms. Therefore, the meaning of the important (from Hegel's point of view) concept of *Aufheben* is clearly triadic as well. In the context of practical realism, this means that Kantianism should not be rejected altogether, but rather carried forward on a different level. It should also be recalled that Hegel's dialectical triad can be understood as a development of Kant's thesis-antithesis mode of thinking in his theoretical philosophy. The model image of the dialectical development is the spiral. Everything tends to repeat itself but on a different level. Thus, we should expect Kantianism to be present in practical realism, but in its rearranged form. How do we understand this? We must start to develop an answer from specifying Vihalemm's conception of *φ*-science.

Vihalemm divided human cognition that science is based on into two types: constructive-hypothetico-deductive and classifying-historico-descriptive (see, for instance Vihalemm,

2001). The former is the basis for the model of φ-science. The constructive-hypothetico-deductive sciences, according to Vihalemm, are physics and the part of chemistry that deals with the laws of nature. We can also use the term 'physics-like' science here. However, the important point is that physics is not the same as φ-science. The latter is a theoretical model of science proper that remains. Physics is developing and at some point, may cease to be a pure φ-science. Thus, the ideal model is necessary.

Rein Vihalemm explained the mathematical content in φ-science as follows: "The foundation of *φ-science* is the mathematical projection operating with experimentally substantiated idealizations, and used to examine the world to identify laws of nature" (2013, p. 8).

In order to make sense of how Kant's approach can be sublated in practical realism, we must pay attention to the meaning of being 'an experimentally substantiated idealization'. This is how Rein Vihalemm describes a scientific object. His further explanation: "It is a certain abstract object – in the case of theoretical concept, an idealized construction – that is a model of a certain actual system given through practice. The meaning (sense) of the concept lies in the referring activity, in the procedures by which the reference is found and the relevant abstract object constructed. It involves the understanding of the role of certain real objects and phenomena, their relations, functions, characteristics, the ways they are handled in a certain system of human activity, i.e. in scientific practice. The concept is like a scheme for constructing, creating or recreating something" (Vihalemm, 2013, p. 8). Thus, the approach is not merely hypothetico-deductive, but specifically constructive-hypothetico-deductive. It must also be acknowledged that the construction of the scientific object is not unique. The nature of the experiment keeps changing (Vihalemm, 2012).

We can see the shift from transcendental idealism and apriorism in the Kantian sense to the practice-based philosophy of science here. The scientist constructs an idealized object for herself to study, but she does this in constant correspondence with objective material reality. Transcendental idealism seems to become sublated – but apriorism not necessarily. It is retained, at least partly, as we shall shortly see.

To be more precise, transcendental idealism may actually be absent altogether from Vihalemm's approach. However, in order to arrive at a better-justified position, let us first consider what transcendental idealism is. Immanuel Kant himself puts it this way: "I understand by the *transcendental idealism* of all appearances [*Erscheinungen*] the doctrine that they are all together to be regarded as mere representations and not as things in themselves [*nicht als Dinge an sich selbst ansehen*], and accordingly that space and time are only sensible forms of our intuition, but not determinations given for themselves or conditions of objects as things in themselves [*als Dinge an sich selbst*]" (A369; the *Critique* is quoted from the Guyer & Wood translation (1998)). There is no *Ding-an-sich* in practical realism. Reality is accessible but not at once in its entirety. The latter would mean assuming the God's-eye view. As we can see, transcendental idealism emphasises the idea of representations that are not things in themselves. The representations are based on human intuition. According to practical realism, the researcher represents what she learns about reality. Space and time really exist. However, the human researcher is still bound to her cognitive capacity. We

can never entirely escape the spatiotemporal nature of our perception. Moreover, we cannot construct the object of research freely or however we wish. Reality imposes constraints, as Vihalemm acknowledges, but so too do the limits of human cognitive capacity. This kind of Kantian apriorism still remains in practical realism – the more so because of the common or at least close positioning of mathematics by Kant and in the conception of φ-science that is the basis for practical realism. However, although Kant sought to establish a solid foundation for metaphysics as a science on the basis of his apriorism, this does not necessarily mean that the Kantian elements of practical realism introduce metaphysics into the approach. After all, what remains of Kant in practical realism can be understood as a sublated form of apriorism.

4. Practical Realism and Metaphysics

Reading the tenets of practical realism (see above) leaves us with the impression that one of the major aspirations of Rein Vihalemm was to propose an approach to science that is free of any metaphysics, i.e. any philosophy whatsoever. Vihalemm had already declared his commitment to a naturalistic approach – a philosophy of science without philosophy – as early as 1993. However, it may not be so easy to dispense with metaphysics.

Before getting to metaphysics directly, let us deal with the God's-eye view issue, the question of whether it is possible for us humans to know everything. Both practical realism and Putnam's internal realism have denied the God's-eye view, as they seek to distance themselves from metaphysical realism. The latter is known for adopting the God's-eye view. Putnam considered such an approach 'realism gone mad', and suggested that his own approach was so 'normal' that it could be called 'commonsense realism', or even just 'realism', as no other approach would adhere to the term strongly enough.

Internal realism had already been criticized before Vihalemm took up the issue, so the criticism itself is nothing new. Nevertheless, Vihalemm quite explicitly identifies internal realism as an unacceptable approach (see tenet 2 above). Vihalemm says the same about "radical" social constructivism. Did he see a link between the two approaches? Hardly so. The reasons for rejecting the two approaches were probably quite different. Vihalemm has characterized social constructivism as self-refuting, arguing that not everything can be constructed because reality itself imposes constraints. But do social constructivists actually claim that anything can be constructed? If they acknowledge the constraining role of reality, then their position may ultimately come quite close to realism.

Rein Vihalemm has claimed that internal realism is not a realist approach altogether. Vihalemm himself puts it like this:

> [P]utnam's 'internal realism', which he also called 'pragmatic realism' belongs, however, to the tradition of Kantianism; that is, it is actually not realism at all (and above I have excluded Putnam's 'internal realism' from the 'practical realism' as I see it) as it scarcely succeeds in avoiding conceptual idealism without a rational reconstruction (Vihalemm, 2011, p. 49).

Still, the problem of difference between practical and internal realism is a tricky one. Internal realism retains strong Kantian flavour, as there is the idea of the correspondence relation between language and some mind-independent entities. This correspondence can hardly be interpreted any other way than through the context of Kantian apriorism. However, the explanation of how this correspondence develops is missing, i.e. the idea of practice is not present in internal realism. This is considered the main difference between practical realism and internal realism, leaving the latter looking considerably less realist. Put differently, in practical realism, apriorism is sublated, whereas this is not the case in internal realism. However, if we consider the five tenets that present the basic framework of practical realism, a similar issue may still arise. Practical realism may not look realist as well, at least in a typical way. The issue of constructive empiricism may appear.

Practical realism explains the correspondence between human cognition and reality by being engaged in connection through practice (see Joseph Rouse above). However, the question remains what we can possibly cognize in the process of practice. Kantian spatiotemporal apriorism still remains. More than that, at times it looks like apriorism in its pure, not in the sublated form. Practical realism is embedded in Kant's apriorism as much as internal realism is from this point of view. After all, the difference between the two may well be just a declarative one.

There is still the issue of truth. The quest for objective truth is an important component of internal realism. The core idea in practical realism is getting rid of the unnecessary quest for truth and taking up the deflationary approach to the issue. Again, Joseph Rouse's approach is important to Rein Vihalemm:

> [t]he predicate 'true' can be applied only to sentences in language. [...] [A language] connects assertions with truth conditions but does not determine whether those conditions obtain. Similarly, what exists depends upon the field of meaningful interaction and interpretation within which things can be encountered. This configuration of practices (including, of course, linguistic practice) allows things to show themselves as they are in a variety of respects. [...] The recognition that the possible ways a thing can be depends upon the configuration of practices within which they become manifest should therefore not reinforce the realist's fear that we are being described as 'world makers'. The language we speak does not determine which of its sentences are true. The practices that constitute our 'world' likewise do not determine which things exist, with what properties (Rouse, 1987, pp. 160–161).

The influence of Rouse is evident when Vihalemm declares: "I cannot speak for pragmatists, but in practical realism, 'truth' can be interpreted in a deflationary way and this interpretation is compatible with semantic realism" (Vihalemm, 2011, p. 55). However, it is difficult, if not impossible to make sense of the progress of science while neglecting the idea of truth as correspondence of scientific statements to reality. This move might shift us away from realism altogether, amplifying the critical voices arguing that practical realism is not really a realist approach, but rather a kind of constructive empiricism. This outcome, however, would introduce a further complication, since Rein Vihalemm has characterized constructivism, at least in its radical form, as self-contradictory. As a matter of fact, this contradiction may provide a safe haven for

practical realism as realism. We cannot construct just anything we like. Reality will resist. But why not retain the correspondence theory of truth rather than seek refuge in a deflationary account? However, since truth is not our primary concern here, we reserve further consideration of this issue for forthcoming papers.

Returning to metaphysics, our main focus in this section, there is a tension between practical realism and Nicholas Maxwell's aim-oriented empiricism (AOE). Rein Vihalemm has declared that Maxwell's criticism of standard empiricism (SE), the current regular understanding of science, meets with approval in practical realism (Vihalemm, 2011, p. 57). However, the essence of Maxwell's criticism of SE concerns exactly the negligence of metaphysical assumptions in science. We have a real problem here. How to understand the statement of Rein Vihalemm that the criticism of SE can be met with approval in practical realism? The essence of Maxwell's criticism of SE is the lack of acknowledgement of the metaphysical assumptions in science by the scientists. These assumptions would be, for instance, comprehensibility of the universe, preferring simple theories to more complicated ones and unified theories to disunified ones. Maxwell brings numerous examples in support of his argument in his many books and papers (see, for instance, Maxwell, 1974, 1998, 2001, 2004, 2007, 2017). There are also treatments of a more general view on Maxwell's thoughts (see, for instance, Maxwell, 2011, 2013). Practical realism, however, is directed at an outspoken denial of anything metaphysical in science.

We cannot find a direct solution to the problem of metaphysics from the works of Vihalemm in addition to his general tendency of giving preference to the naturalist approach of studying science without philosophy. Still, we can do guesswork, adhering to Vihalemm's typical approach to science, which he reached in his last papers. We can also hope for getting help from the authors Vihalemm has built on. First of all, the hope is with Joseph Rouse at this point again. Both Vihalemm and Rouse were interested in the practical interaction of the researcher with the world. The philosophical conceptualization of this interaction is not important; what matters is scientific practice itself. The metaphysical assumptions of Maxwell are there as inherent components of science. However, the three assumptions Maxwell keeps emphasising need not be viewed as metaphysical principles, but rather as useful methodological rules scientists need to follow.

There appears to be room here for an analogy between Vihalemm's understanding of metaphysics and his understanding of truth. In a sense, his attitude toward metaphysics might also be described as deflationary, particularly in light of his response to Nicholas Maxwell's criticism of standard empiricism. Indeed, the term "deflationary metaphysics" is already in use. We therefore need to consider whether Vihalemm's possible understanding of metaphysics fits the sense in which the term is used, for instance, by Sergio Sismondo, or whether "deflationary metaphysics" can be understood in different ways. Rein Vihalemm himself has not used the term "deflationary" in relation to metaphysics.

Sismondo connects the deflationist turn in philosophy of science to Arthur Fine's "Natural Ontological Attitude" (NOA) (Fine, 1986a, b), an approach that was also taken up by Joseph Rouse (1991) (Sismondo, 1997, p. 219). According to Sismondo, NOA is

"a minimalist philosophical position, trying to formulate an overlap of realism and instrumentalism as applied to science, where instrumentalism is taken as representative of antirealism more generally, including positivism, pragmatism, and constructivism" (Sismondo, 1997, p. 219).

This explanation is complicated but gives quite a good idea about the essence of NOA. The idea of overarching realism and instrumentalism alone is intriguing. It is of special interest to us that instrumentalism is taken as a general representation of anti-realism here. Rein Vihalemm set the same task for himself with practical realism, namely providing an alternative approach to both realism and anti-realism. In addition, Vihalemm's understanding of science is naturalist as well. As we know, he has advocated for this approach since the early 1990s.

It appears that Sismondo is also looking at science as a practical activity or at least considering practice an important aspect of science. Interestingly, it is just from the point of view of science as practice that Sismondo finds it difficult to avoid metaphysics. Sismondo explains: "When we investigate scientific research from the bottom up, when we investigate practice, we can be led very quickly to the type of metaphysical questions that deflationists like Fine are trying to avoid. The "core position" that we can accept many scientific beliefs as true may be right, but to make it right demands metaphysical work" (Sismondo, 1997, p. 221). This is simpler than many of us would guess. We accept some scientific beliefs as true, not necessarily considering upon which grounds we can do so. This is where metaphysics comes in. We need the support of metaphysics in order to take the positions about scientific claims. As the influential Hungarian Marxist thinker György Lukács noted, in a passage appropriately taken up by Sismondo, specific scientific practices "aim at truthful reflection of reality" (Lukács, p. 23). It is not entirely clear, however, whether this commitment should be understood as an inherent component of science, as Maxwell claims, or not. Still, it very much looks like Maxwell would say that the truthful reflection of reality would be impossible to achieve without the basic metaphysical assumptions in science.

Sergio Sismondo specifies: "Deflationary metaphysics creates contexts for seeing worlds as constructed, actions as pragmatic, or scientific descriptions as realistic" (1997, p. 230). It is research practice that is important. Not any kind of general considerations. There is a clear fit with practical realism if we change the worlds being constructed to objects of science being constructed. In practical realism, actions are understood in practical (or pragmatic) terms, while scientific descriptions are understood in realist terms. The deflationary mood about laboratory and theoretical practices tells us, according to Sismondo, that the knowledge scientists produce is not simply about the natural world – at least not about the untouched empirically given world (1997, p. 230). Of course not. This is one of the central points of practical realism. The world, as it is accessible to the scientist, is not free of norms. Human practical activity, like scientific research, however, has its normative side. This may mean that, in such a case, realism itself becomes problematic, while the possibilities for metaphysical interpretation remain open. It is also Sismondo's final claim that scientific practice does not remove all scope for metaphysical interpretation (1997, p. 230). Sure, but do we need these interpretations? And does realism really become

problematic? From the point of view of both Vihalemm and Rouse, the researcher is necessarily in contact with reality in the course of her activity. Why take any special positions about whether any scientific claims are true or right? If we can establish contact with reality, then our methods should be effective, since successful interaction with reality suggests that they are working properly. There is one reality. That one reality is the only true one. Still, our interpretations of research results concerning that one and only reality may be inadequate. Reality is just there. In principle, truth claims and metaphysical claims about it can really be treated as deflationary. However, in the case of truth, this is natural and acceptable, while in the case of metaphysics this is not so obvious. In the case of truth, we can say that truth is a property of propositions, whereas the world simply is as it is – it is neither true nor false. The case of metaphysics, however, is not so straightforward. In order to convey a consistent scientific world picture, we have to make claims that are not verifiable or falsifiable. We may call them methodological rules rather than metaphysical statements as the critics of Nicholas Maxwell's approach claim, but they are there. The results of practical research would be floating around like the birds of knowledge in Plato's aviary without them.

It may be that we have, in fact, simply presented another way of understanding deflationary metaphysics. The latter has its place in science in the form of nontestable statements that can be called methodological rules if we wish. Metaphysics may have become obsolete, perhaps even exhausted, but simply denying it does little to advance our understanding of either science or reality. Instead, we need to develop a clearer account of what it means to understand metaphysics in science in deflationary terms.

5. Conclusion

Despite all the efforts of the naturalists and adherents of the practical approach in the philosophy of science, metaphysics is not fully exhausted and can hardly become such. Just the opposite is true. The possibility, and perhaps even the need, for metaphysical interpretations continues to arise in different forms across different approaches to reality, even though the reality they seek to understand is one and the same. Scientists may not need metaphysics as some kind of external foundation for their field. However, this does not mean that metaphysics disappears from the picture. Rather, it may emerge directly within the scientific picture itself, either as an inherent component of science, as Nicholas Maxwell has argued, or perhaps in a deflationary form.

Despite Rein Vihalemm's claims to have sublated transcendental idealism and rejected metaphysics, Kantianism remains visible in practical realism. Kant's conception of *eigentliche Wissenschaft* clearly provides the basis for Vihalemm's φ-science, a connection that Vihalemm himself also acknowledges. The latter, on the other hand, serves as the backbone of the practical realist approach to science. Still, practical realism cannot be taken as just a development of Kant's ideas or any kind of pre-existing realism. It is a very interesting original contribution. Practical realism needs critical analysis and development. At some point, may be even sublating it would become the

preferable strategy, because we cannot presume that the best interpretation of scientific research has been achieved under its roof. Not just the theories and methods of science have to be assessed critically constantly but our aims as well. The latter task, however, cannot be just theoretical. Science is a theory-guided practical activity that puts us into contact with reality in a specific (scientific) way. However, we must continue to analyze and explain the scientific way of understanding reality, precisely because our understanding of science itself can never be regarded as complete.

Acknowledgement: Completing this paper was supported by the Estonian Research Council Grant PRG462.

References

Fine, A. (1986a). *The Shaky Game: Einstein, Realism and the Quantum Theory*. Chicago: University of Chicago Press.

Fine, A. (1986b). Unnatural Attitudes: Realist and Instrumentalist Attachments to Science. *Mind*, 95, 149–179.

Glazebrook, T. (2001). Heidegger and Scientific Realism. *Continental Philosophy Review*, 34, 361–401.

Heidegger, M. (1977). The Age of the World Picture. *The Question Concerning Technology and Other Essays*. (pp. 115–154). New York, London, Toronto, Sydney: Harper Perennial.

Heidegger, M. (1989). Küsimus tehnika järele. *Akadeemia*, 6, 1195–1228.

Heidegger, M. (2011). [1964] The End of Philosophy and the Task of Thinking. In D. F. Krell (Ed.) Martin Heidegger: Basic Writings from Being and Time (1927) to Task of Thinking (1964). With a foreword by Taylor Carman. London & New York: Routledge Classics, (pp. 311–325). [Original: Heidegger, M. (1964) Das Ende der Philosophie und die Aufgabe des Denkens. Essay written for the UNESCO colloquium "Kierkegaard vivant", Paris, April 21–23].

Kant, I. (1998). *Critique of Pure Reason*. Cambridge: Cambridge University Press.

Kant, I. (2002a) [1783]. *Prolegomena to Any Future Metaphysics that will be able to come forward as Science*, (G. Hatfield, H. Allison & P. Heath, Trans.), in P. Guyer & A. Wood (Eds.) Theoretical Philosophy after 1781. The Cambridge Edition of the Works of Immanuel Kant in Translation, Cambridge: Cambridge University Press [Original: Kant, I. (1783), Prolegomena zu einer jeden künftigen Metaphysik, die als Wissenschaft wird auftreten können]. Citations are given, as usual, with volume and page numbers of Kants Gesammelte Schriften.

Kant, I. (2002b) [1786]. *Metaphysical Foundations of Natural Science*, (M. Friedman, H. Allison & P. Heath, Trans.) in P. Guyer & A. Wood (Eds.) Theoretical Philosophy after 1781. The Cambridge Edition of the Works of Immanuel Kant in Translation, Cambridge: Cambridge University Press. [Original: Kant, I. (1786). Metaphysische Anfangsgründe der Naturwissenschaft]. Citations are given, as usual, with volume and page numbers of Kants Gesammelte Schriften.

Kochan, J. (2011). Getting Real with Rouse and Heidegger. *Perspectives on Science*, 19(1), 81–115.

Lukács, G. (1963). *The Meaning of Contemporary Realism*. (J. and N. Mander, Trans.). London: Merlin Press.

Lõhkivi, E. & Vihalemm, R. (2012). Philosophy of science and practical realism. *Studia Philosofica Estonica* (Guest Editorial) 5(2).

Marx, K. (1845). *Theses on Feuerbach*, Retrieved April 13, 2024, from http://www.marxists.org/archive/marx/works/1845/theses/index.htm

Maxwell, N. (1974). The Rationality of Scientific Discovery. *Philosophy of Science*, 41(2), 123–153.

Maxwell, N. (1998). *The Comprehensibility of the Universe: A New Conception of Science*. Oxford: Oxford University Press.

Maxwell, N. (2001). *The Human World in the Physical Universe: Consciousness, Free Will, and Evolution*. Lanham: Rowman & Littlefield Publishers, Inc.

Maxwell, N. (2004). *Is Science Neurotic?* London: Imperial College Press.

Maxwell, N. (2007). *From Knowledge to Wisdom: A Revolution for Science and the Humanities*. 2nd Edition. London: Pentire Press.

Maxwell, N. (2017). Understanding Scientific Progress. Aim-Oriented Empiricism. *Modern Physics*, 35, 109–119.

Mets, A. (2019). A Philosophical Critique of the Distinction of Representational and Pragmatic Measurements on the Example of the Periodic System of Chemical Elements. *Foundations of Science*, 24, 73–93.

Mets, A. (2024a). What does '*φ*-Scientificity' Mean? I. Models and Measurement: Galileo. *Acta Baltica Historiae et Philosophiae Scientiarum*, 12 (1), 6–36.

Mets, A. (2024b). What does '*φ*-Scientificity' Mean? II. Measure, *Acta Baltica Historiae et Philosophiae Scientiarum*, 12(2), 3–31.

Mets, A. (2025). What does '*φ*-Scientificity' Mean? III. Material Quantities. *Acta Baltica Historiae et Philosophiae Scientiarum*, 13(1), 3–30.

Müürsepp, P. (2011). Knowledge in Science and Non-Science. *Baltic Journal of European Studies*, 1, 61–73.

Müürsepp, P. (2012). The Changing Role of Scientific Experiment. *Studia Philosophica Estonica*. 5.2, 152–166.

Müürsepp, P. (2013). The Aim of Science – Knowledge or Wisdom. *Problemos*. 84, 72–83.

Rouse, J. (1987). *Knowledge and Power: Toward a Political Philosophy of Science*. Ithaca: Cornell University Press.

Rouse, J. (1991). The Politics of Postmodern Philosophy of Science. *Philosophy of Science*, 58, 607–627.

Sismondo, S. (1997). Deflationary Metaphysics and the Construction of Laboratory Mice. *Metaphilosophy*, 28(3).

Vihalemm, R. (1993). Teadusfilosoofia ilma filosoofiata? (Summary: Philosophy of Science without Philosophy?). *Akadeemia*

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 271 – 282

THE INFLUENCE OF 'PHILOSOPHY OF THE ILLUMINATED' IN THE KHOMEINI'S APPROACH TO CREATE AN 'IDEAL SOCIETY'

Bogdana Todorova*

Abstract: The study examines how the philosophical ideas of the so-called 'Illuminated' are reflected in Khomeini's vision of an 'ideal society'. It can be argued that the Islamic mystical tradition (thereafter Irfan) to which he was a devotee played an important role in the formation of his worldview in general, particularly the role of the mysticism and his responsibility for orienting people towards the sacred realm through the establishment of an ideal society. For the purposes of the study, alongside the historical approach, an anthropological approach has been applied, which is more open to qualitative research. The results suggest that his later political theory of *Wilayat al-Faqih* can be legitimized on the basis of gnostic and metaphysical notions, an issue that has important implications for future research. Using Plato's and al-Fārābī's views of "ideal society" as a motivational and inspirational power he created a new "semantic power", as a basic indicator for understanding the principle of development of society. The study aims to stimulate further debate within the social sciences on the relationship between mysticism, metaphysics, and political thought.

Keywords: Ideal society, Ayatollah Khomeini, Philosophy of Illuminated, semantic power.

1. Introduction

Religious movements in the mid-1970s sparked research interest in non-Western traditions and cultures, stimulating new theoretical explorations. Another argument underlying this research is that religion remains a motivating force in Iranian politics. The rise of Shi'i Islam as a dominant ideology is rooted in the institutional development of the *ulamā* (scholars of Islamic doctrine and law) in the nineteenth century in Iran. However, it was reproduced by diverse Islamic intellectuals such as Ayatollah Khomeini, Ale Ahmad and Ali Shari'ati (Nasr, 2006). It may be difficult to understand the strategy of a state if one has not studied the philosophy of those who govern it (Lubna, 2008, p. 11).

The teachings of Islam are closely related to the concept of community. Khomeini was deeply attached to the Islamic mystical tradition, more commonly known as Sufism. The term that came to be widely used to refer to Shiite gnosis is *Irfan*. Sufi concepts such as the *Qutb* (the spiritual Pole of the universe) came to be associated with the Shiite Imams, as did the notion of *Wilayat*, which represented the continuation of divine grace and guidance through the Imams after divine revelation ceased with the passing of the Prophet (Bazzi, 2023). As a religious leader, Khomeini sought to incorporate this form of Sufism without compromising the tenets of the Qur'an and

*Institute for Philosophy and Sociology, Bulgarian Academy of Sciences; bonytodorova@gmail.com

©2026 Bogdana Todorova. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

Sunnah in his pursuit of an ideal society. He is a follower of the hidden imam, whom Shiite Muslims believe is in hiding and will reappear at the end of time to establish a just order in the world. While there was general agreement among Shiite jurists that the expert on divine law enjoyed a measure of authority in the age of the occultation (Sachedina, 1998), it was Khomeini who declared that this authority should extend to all affairs with which the hidden Imam himself was charged, including the setting up of a divine order on earth (Khomeini, 1983). This outlook informed his political worldview and helps for the formation of his most important political theory *Wilayat al-Faqih* (Guardianship of the Jurist). The ontological background which influenced Khomeini's worldview in general serves to shape his Gnostic worldview and the role of the *mystic wayfarer*, both spiritual and social, in it.

According to Sayyid Hossein Nasr this proved the strong relationship between Sufism and Shiism (Nasr, 1972, p. 105), whereas for Corbin, this process was simply a matter of the Shiites reclaiming their own Sufism, which he regarded as Shiite in origin and significance. (Corbin, 1971, p. 219). We can find its culmination in the works of Mullā Ṣadrā (c. 1571/2 – c. 1635/40 CE / 980 – 1050 AH), who introduced a theosophical system he named *Transcendent Theosophy*, where he brought harmony between the Qur'anic teachings, the heritage of Greek philosophy and Sufi Gnosis, all within the parameters of Shiite theology. Such a relationship was analyzed by Lloyd Ridgeon, who described the person of Khomeini as "hidden" (Ridgeon, 2014). In connection with this, several studies have specifically examined the role of the mystic as a political leader (Bazzi, 2023; Fehri, 1982; Sachedina, 1998).

The article brings into relief elements of Khomeini's reformist agenda and emphasizes the impact of his interpretive approach on mystical tradition and reformist thought in the Muslim context. The link between the philosophy of illumination and the Islamic concept of community in creating an ideal society is examined through a philosophical approach, which underscores the novelty of the research. Khomeini's understanding of *Irfan*, more precisely his gnoseological views, can be seen as a motivational and inspirational power for his political theory. The result is a new "semantic power", serving as a basic indicator for understanding the development of society. Semantics play an important role in the nominative process, depicting the features in the human world, the specifics of the cultural and historical experience of the people (Arzhanovskaya et al., 2018). The concept of power is analyzed within the context of legal institutional discourse, distinguishing this research from other studies that focus primarily on how the Iranian Revolution shaped a new political order through Khomeini's ambitious goal of reviving and spreading Shia Islam.

2. Methodology and conceptual framework

To examine the influence of the so-called "philosophy of illumination" on Imam Khomeini's political theory, the study employs a descriptive qualitative approach aimed at understanding the various concepts identified during the research process through content analysis and library research. As a research method, content analysis provides a systematic and objective means of describing and quantifying phenomena (Schreier,

2012). For the purposes of analysis, the study data are reduced to concepts that describe the research phenomenon (Elo & Kyngäs, 2008).

The research draws on secondary sources, including previous research, scholarly articles, and reference books addressing topics related to the study. It also makes use of materials from the Imam Reza Library in Mashhad and the Marashi Library in Qom, established by Shahab ad-Din Mar'ashi Najafi, one of the five Grand Ayatollahs at the time of the Iranian Revolution, who envisioned Qom as a center for the education of Shiite clerics and the formation of Shiite political and administrative elites.

2.1. The religious seminary in Qum and the influence of the "Illuminated"

Qum is the place to educate and prepare the spiritual leaders of the country. Seminarists are associated with different traditions, theologians and lecturers, some of whom are of Iraqi origin, selecting Qum to avoid persecution in their native land. These are the people who helped Tehran to quickly build relations with Lebanese Shiites in 1980, who are closely related with another religious center – Najaf. During elections, the place is visited by the political leaders of Iran, who want to pay tribute to their own spiritual mentors and to receive their blessing. The challenges of modernity have their reflection in the two spiritual centers and stimulate lecturers to go beyond the limits of traditional sciences in the instruction of new disciplines. Today, Qum is an educational center for the new generation of Shiite politicians and religious leaders not only for Iran but also for the entire Middle East and South Asia.

The arrival in Qum of Ayatollah Mohammad Ali Shahabadi, who became Khomeini's most important teacher in *Irfan*, began a seven-year period in which he embarked on an intense cultivation of both the theoretical and practical aspects of the discipline. He was introduced to Manazil al-Sareen (Stations of the Wayfarers), by Shahabadi, "to whom [his] spiritual life is indebted" (Khomeini, 1990, p. 78).

The result of this mentorship included the two most important books Khomeini authored on *Irfan*, which were to shape his Gnostic worldview and the role of "the mystic wayfarer", both spiritual and social, in it. These were respectively *Misbah al-Hedaya ila al-Khilafah wa al-Wilayah* [hereafter Lantern of Guidance]), and *Taaliqat ala Sharh Fusus al-Hikam wa Misbah al-Uns* [hereafter Glosses]). Both of these books betray Khomeini's indebtedness to the cosmogony of Ibn 'Arabi (Chittick, 1989). Central to that worldview is the theory of Emanation according to which God, the Absolute, makes Himself known through a process of self-manifestation culminating in the appearance of the cosmic world and all creatures in it.

Khomeini shows a remarkable interest in disciplines such as philosophy (under the influence of the famous peripatetician and representative of the so-called "philosophy of illuminated" (Sayed Qazwini) and Gnosticism (*Irfan*) (a less popular topic of education in madrasahs at that time), as well as a special interest to the translation of Qur'an of Mullā Muhsin Fayd Kāshānī, study the *Fusus al Hikam* of Ibn Arabi and Tafsir and Safi. Khomeini's lecturer in ethics was Jawad al-Maliki Tabrizi, whose influence can later be seen in the young Khomeini's course on the "Bases of Ethics."

Addressing contemporary problems, the course became so popular that it attracted not only theology students but also ordinary citizens from Qom, Tehran, and even distant Isfahan, at a time when the government was attempting to restrict the influence of the ulamā outside religious institutions.

The theme of gnosis as a form of knowledge was central to Khomeini's intellectual development. In a short autobiography written in 1934, he acknowledged the profound influence of Mullā Ṣadrā (d. 1050/1641), particularly his *Hikmat Al Muta'alyah fi-l-asfar al-'aqliyya al-arba'a* [The Transcendent Theosophy in the Four Journeys of the Intellect]. This work constitutes a philosophical encyclopedia addressing major questions in Islamic philosophy, drawing on the ideas of earlier thinkers from Pythagoras to philosophers contemporary with Mullā Ṣadrā and developing responses to them through new arguments. Khomeini was also influenced by *Iqaz al-na'imin* [Awakening the Asleep Ones]. These works became an integral part of his spiritual outlook, which was subsequently reflected in his political activity. Khomeini understood himself as following this spiritual path, establishing a relationship with God and His attributes and, on this basis, assuming the capacity to act in the human realm as a divine lawmaker.

Khomeini was among those who understood how Mullā Ṣadrā could be at once a Peripatetic philosopher, an Illuminationist sage, and a Muslim mutakallim, as well as what distinguished him from other thinkers. Mullā Ṣadrā transformed not only the framework of philosophy but also sought to free human thought and intellectual inquiry from unnecessary difficulties. He also spent a period of time in a village near Qom, where he wrote several works, including the *Risāla fi 'l-ḥashr* and the *Risāla fī ḥudūth al-'ālam*. His thought engaged with principles of the earlier Illuminationist tradition, including the primacy of existence, the multidimensionality of light and existence, the trans-substantial motion of nature (ideas associated with Persian and pre-Socratic sages), the union of intellect and intelligible, and the principle of "the most simple reality," also found in the mysticism of Plotinus and Sufism. The simplest reality pervades in everything, while being none of them. He adduced rational arguments of them and put them in the guise of demonstrative philosophy. The great philosopher introduced a new method of philosophical inquiry that changed the course of philosophy and significantly enhanced its intellectual standing. Mullā Ṣadrā regarded philosophical thought as an integrated whole, holding that reason, revelation, and intuition ultimately derive from the same source. From his perspective, truths apprehended through intuition or knowledge by presence, insofar as such intuition is genuine and divine, must also be susceptible to rational demonstration. His philosophical method therefore sought to incorporate and unify different modes of attaining knowledge. For this reason, Mullā Ṣadrā did not characterize his school as exclusively Peripatetic and demonstrative or as simply Sufi (The Metaphysics of Mullā Ṣadrā, also known as *Kitāb al-Mashā'ir*) (Ṣadr al-Dīn Shīrāzī, 1964).

He was also well acquainted with the doctrines of the *mutakallimūn*, although he was critical of both Mu'tazilite and Ash'arite approaches. In his view, the Mu'tazilites had gone astray by following what he regarded as the mirages of Peripatetic philosophy, while the Ash'arites, in their effort to safeguard the *shar'*, had distanced

themselves excessively from reason. For Mullā Ṣadrā, however, the *shar'* and reason were complementary: when joined together, they constituted "light upon light." He thus provided philosophy with the conceptual tools necessary to defend Islam while establishing a firm foundation for the relationship between reason and revelation.

By bringing *sharī'ah* and rational argumentation into a mutually supportive relationship, Mullā Ṣadrā incorporated both within a unified philosophical vision. He maintained that there could be no fundamental inconsistency between the teachings of the prophets and revelation, on the one hand, and truths attained through the intellect, on the other. Accordingly, he affirmed both reason and rational demonstration as well as Illuminationist wisdom and intuition. Likewise, he accepted both *ta'wīl*, or hermeneutic interpretation of the Qur'an, and interpretation concerned with its outward verbal meaning. His Transcendent Theosophy therefore sought to encompass and move beyond the limitations of both Sufi epistemology and conventional theology.

Within this framework, reason serves as a means of understanding the *sharī'ah* and guiding human beings toward knowledge of the truths of the world; properly exercised, it cannot conflict with revelation. Reason and revelation ultimately proceed from the same source – the Holy Spirit or Active Intellect. Reason, when free from corrupting influences, acquires knowledge through its connection with the Universal or Active Intellect, while revelation reaches the prophet through this same intelligible source. Intuition likewise constitutes a reliable means of apprehending truth. Philosophical knowledge consequently involves a form of experience as well as intellectual demonstration, since Illuminationist wisdom incorporates both practical and theoretical wisdom. Through the integration of these different modes of knowledge, Mullā Ṣadrā establishes a unified epistemological circle in which reason, revelation, and intuition mutually reinforce one another.

2.2. Khomeini's understanding of Illuminationist wisdom

Following the later development of Mullā Ṣadrā's thought in particular, Khomeini demonstrates that the school of Transcendent Theosophy is not a static philosophical tradition. Drawing on his own spiritual development, he engaged deeply with the Qur'an and sought to analyze philosophical questions through the perspective of Qur'anic verses. Like Mullā Ṣadrā, Khomeini also expanded the aims of philosophy. He connects politics, traditionally regarded as the culmination of practical wisdom and concerned with the proper administration of society and human happiness, with a broader conception of practical wisdom directed toward the attainment of both worldly and ultimate happiness.

Within this framework, resurrection is associated with practical wisdom, moral transformation, and the purification of the soul. It represents the return and culmination of the human being's progressive movement toward the "origin of emanation" and existence (Rahman, 1987, p. 234). Khomeini thus sought to present this movement toward the Divine as an ultimate goal of philosophy. Building upon Transcendent Theosophy, he points toward a wisdom that extends even beyond the "transcendent," toward the "Throne" of the pure Divine realm.

According to Shi'i thought, Divine acts, like all acts of wisdom, are directed toward a purpose. This purpose, however, cannot consist in God's attainment of perfection or avoidance of deficiency, since neither can be attributed to the Divine Essence. Rather, it concerns the eternal and everlasting emanation of God and, from the perspective of Divine knowledge, the good of humanity. This principle is expressed in the wisdom of faith associated with Mīr Dāmād and in Mullā Ṣadrā's Transcendent Theosophy (Khamenei, 2000).

Reason, in this account, functions as a mediator that guides the human being along the path toward the Divine. Contemplative concentration on the Divine Presence seeks a form of identification with Divine knowledge, which is understood as the external manifestation of the light of the eternal Imam. Knowledge of God is therefore proportionate to the extent to which the individual opens the soul and heart to the Truth.

Khomeini came to adopt a methodology which incorporates Mullā Ṣadrā's teachings about the four (mystical) travels (the mystical wayfarer undertakes a journey which consists of four mystical travels that culminate in the mystic achieving annihilation, then subsistence by God, and then return to the world to be a guide) within the cosmogony of Ibn Arabi, which places the Perfect Man at the pinnacle of the stations related to proximity to God.

Each stage of manifestation according to Khomeini (Khomeini, 1983, p. 20) represents a *khilafah* (deputy or vicegerent, more commonly associated with the word caliph, or representative of "people" in Islam) of the previous stage. For instance, the second major stage of manifestation, that of al-Asmaa al-Husna (the Beautiful Names of Allah, which include the Merciful, the Knower, the Proud, the Punisher, etc.) is, according to Khomeini, the adept of the initial stage of Ahadiyyah (Unity of God) in which God is an unknown and unknowable mystery. The adept of the Beautiful Names of Allah forms the third stage of manifestation, that of "the Immutable Entities". The last stage represents the emergence of the cosmic world. This process of the self-manifestation of God is called the Arc of Descent.

Khomeini used this kind of knowledge as a motivational and inspirational power to develop his political theory later. What emerges clearly from Khomeini's central gnoseological ideas, is that "Wilaya", which according to him, is associated with nearness to God gained through mystical wayfaring, and forms the esoteric dimension of *khilafah*, implying deputyship and social responsibility. In addition to that, Khomeini saw that the station of the Perfect Man is a hierarchical position with the Prophet and Imams situated at the top, followed by other Gnostics and wayfarers, and indeed pious people in general, each according to his level. For him, the Perfect Man is the person who has accomplished the fourth travel in his mystical journey, and after having achieved annihilation in God and then sobriety through God, "comes back", as it were, in order to guide humanity (Mutahhari, 1983). Indeed, "the true spiritual traveler, after completing his journey towards God and for the sake of God, tries to perfect the servants of God and embarks on reform and development of society" (quoted in Damad, 2000, p. 34).

The relevance of the studied problem is also related to modern linguistic trends – growing attention to the different discursive spheres correlated with communicative fields. Discourse is one of the most complicated scientific concepts incorporating not only linguistic but also extra-linguistic components (Eltanskaya et al., 2018, p. 5). The institutional component of discourse is an integral part of the mechanism of society's development. Institutions make the rules defining at the same time opportunities and constraints, rights and duties, roles and statuses. Changes in the functional forms of social connections and relationships cannot occur without corresponding structural transformations. Together, these functional and structural changes constitute the mechanisms through which societies evolve. Consequently, neglecting the importance of institutional change, or failing to coordinate such change with structural and organizational reforms, can undermine efforts to transform a community or social group and ultimately contribute to their failure (Eltanskaya, 2011). Khomeini employs status-oriented political discourse as a form of institutional communication involving interaction between representatives of a particular group, namely the Imams, and members of society. Within this framework, individuals exercise the opportunities associated with their respective social statuses and roles through existing public institutions, whose number and functions are shaped by the needs of society at a particular stage of its development. To determine the type of institutional communication, it is necessary to consider the status-role characteristics of the participants of communication (judge – lawyer) and the role of legal discourse – making rules and regulations (improvement of legislation) (Eltanskaya, 2011). These characteristics should justify his rules and regulations, which should facilitate their functioning in an easily recognizable, legalized form over a long period of time. Thus, he creates a new "semantic power", which serves as a basic indicator for understanding the development of the society. Using it as a tool, Khomeini is able to understand the truth, with its spiritual power and present it to the others. The concept of power is analyzed within legal institutional discourse because law is an area in which the dynamics of social change, shifts in public consciousness, and the mentality of the leader-jurist are clearly manifested (Sheigal, 2004). The latter has both social and linguistic conditionality. The Shi'a "rational-epistemic" scheme used reason to extend the law to the new territories.

Khomeini's legal lexicon explains the relevance in terms of the theme of the concept of "linguistic identity" (Eltanskaya et al., 2017). Thus, this approach allows us to study a person both as a representative of society and as a unique, inherently valuable individual. In such a paradigm, the person is recognized as a measure of all things. Reality is reduced to the leader's message, to his power and the power of the language he speaks through him, but also within him – an expression of the greatness of an ontotheology, of something that encompasses everything and provides the broadest possible framework for discourse.

Language is the most fundamental element of Iranian identity, one of the leading dimensions in the vision of a civilizational model on which the Iranian state and society must build, alongside the educational, cultural, social, and economic processes directly related to the transition from tradition to modernity. It is charged with the symbolic function of representing what is authentically Iranian.

The main driving force behind the socio-political changes in Iran is semantic power, based of gnostic and metaphysical notions. The records of Khomeini's speeches in Paris during his expulsion cause an emotional response among the exalted crowd in Iran. It is the symbolism of language, in the spirit of Whitehead, that has a figurative and emotional impact (Sycheva, 2000, p. 67). His language is a means of influencing the psychology of the nation, thus emphasizing its social function. Language is specific insofar as in it the symbolic is free – the speaker reproduces it by moving from the object to the words, and the listener from the words to the object (Whitehead, 1999, p. 8). The linguistic dimension is a metaphysical path to accessing the spirituality necessary to achieve an ideal society, and semantic power is the means of building it, a consequence of Islamic mysticism.

2.3. The "Ideal society"

A Muslim as a servant of Allah has social rights and responsibilities to the community. Hence, this would require the formation of *an ideal society* which meets the needs of the individual and society. The *ideal society* as a philosophical concept has origins in Plato's most famous and widely read dialogue, *Republic* (509d–511e). For Plato, the ideal state is as immutable as a Doric temple, since any change can lead only to decadence and corruption. Society must therefore be protected from all that could upset the civic order and induce change. The guardians must devote themselves entirely to the service of the state. They may not possess material riches (which give rise to jealousy and conflict); they may not indulge in frivolities (which could compromise their integrity); nor may they entertain private ambitions (Hummel, 1994).

Abū Naṣr al-Fārābī (872-950) has also highlighted the concept of *al Madinah al Faḍilah* (the virtuous city) (Kamal, 1995). He described it in his best-known work *Mabādi' ārā' Ahl al-madīna al-fāḍila* [The Principles of the Opinions of the People of the Virtuous City]. *Madīna* is a Qur'anic word and means the city in general and the city of the Prophet in particular (*Madīnat al-nabī*). As a reader and interpreter of the Platonic *Republic*, he relied on the Greek concept of the polis (Gutas, 2004). In his *Perfect State*, al-Fārābī explicitly states the needs of *indhār*, a philosopher-prophet (al-Fārābī, 1985). He is the only one who can govern with knowledge of the true purpose of human existence, which is obtained through physical and metaphysical investigation (Walzer, 1967). But if the balance in society collapses, these perfect states will go astray, and people will lose truth, happiness, the right path, morality, stability, solidarity and all virtues which had been achieved under the leadership of the prophet ruler (Akhmetova, 2017).

Khomeini draws on Fārābī's conception of religion as grounded in a perfect demonstrative philosophy. Such a perfect or virtuous religion is based on dialectical or rhetorical premises, or tenets, established by the philosopher, who possesses the capacity both to ascertain the truth and to recognize what genuinely represents it (Netton, 1985).

In the West, scientists have proposed various concepts of the "ideal society", such as socialism and capitalism as models of an ideal society among adherents of Marxism

and individualism/liberalism respectively (Mutahhari, 1986). Against the background of a new phase in the crisis of neoliberal capitalism – a condition to which philosophers and sociologists have applied Gramsci's concept of the "interregnum," a period of indefinite duration in which the old order is breaking down while a new one has yet to emerge, and unpredictable events replace established structures – Khomeini proposed a practical and ethical Islamic reform aimed at creating an ideal society. This reform involved a qualitative transformation from a purely religious Islam toward a cultural-religious Islam, which he regarded as essential to the evolution of Iranian society. His most important aim, grounded in the principle of "Unity – a great Truth," was to construct a theo-democratic model of society.

According to Imam Khomeini, the relation of man to God is divided into two aspects – the aspect of recognition and the practical aspect. The aspect of recognition is the human ambition to know God and His qualities, and the practical experience reflects service and human humility before God. Freedom has to be discussed in the context of human volitional actions. From a theological perspective, efforts to know God are commendable and require purposeful action to understand His attributes, the limits of which are defined through revelation. Knowing God is important and prophets are sent for this reason. This knowledge is precise to such an extent that it needs a number of efforts, although knowledge and acceptance of the existence of the sole God are theoretical acts by their nature. Regarding the practical connection, humility and service are related to three types of relation: the relation of a man to himself, his relations to others, and his relation to nature. Specific elements are blessing and worship, which are defined by dogmas prescribed by the Almighty, and the minimal level of corresponding actions is defined through revelation. A maximal level is understood through performance of the tasks and accomplishment of goals, which are set in complete human life. Thus, an excess in prayer can lead to cessation of performance of public obligations or it can become a reason for physical weakness or disease, or provoke a crisis in the relation of a man to himself and others. According to Khomeini, the world in its completeness provides good conditions for man to gain knowledge and reach worthy goals. The society referenced by Khomeini rests on the purity of spirit and actions and leads to the new approach of practical and ethical Islamic reform and qualitative transformations in the evolution of Iranian society and the Muslim world as a whole, where justice occupies a central place.

Imam Khomeini believes that the term "society" has a full meaning and should stimulate the evolution of certain traditions and customs rather than blocking it. Personal liberty should not endanger the freedom of others, and individual liberty should not harm social liberty. A truly commendable system of human relations will appear as soon as human values become a factor for the estimation of people. He develops the idea for an *ideal society* as a new type of collaborative society based on knowledge and morals. Moral values reflect the relation man – society – nature, and not the interests of a privileged group of individuals. Therefore, justice concerns not only individual interests but also public interest (*maslaha*), which is a key concept in Islam. The subject of jurisprudence was never divorced from other aspects of Islam, especially the mystical one, as far as Khomeini was concerned. According to Sayyid Ahmad Fehri, who was also the editor and translator of many of Khomeini's books on

Gnoseology, Khomeini's approach was to "demonstrate the conformity of the sharia to the logic of Irfan as well as the conformity of Irfan to the logic of the sharia" (Fehri, 1982, p. 11). Islam does not mean refitting theology, but rather reviving the moral compass of the people.

In order to create such a society, the role of educators is vital and cannot be underestimated. Muslims must learn, understand and appreciate the teachings of the Qur'an, and practise it because the Qur'an is regarded as the pinnacle of both spiritual and behavioral guidance for all Muslims. It is a part of life for every individual, the law for political rule and the guidance in preaching to God (Aboul-Enein, 2016; Usman & Kadir, 2020). Islamic teachings guide humans to live well on earth as being an earthly caliph of Allah. Individuals who are able to appreciate and understand the essence of the verses of the Qur'an will realise the importance of establishing a fraternity of ascetics for the sake of harmony and prosperity without creating hostility and posing threats to people's lives. The Prophet Muhammad has managed to develop and build an ideal society through mosques that establish and foster ties between all Muslims.

All human deeds and actions need to be constrained by guidelines and the laws of Allah in order to balance the demands of the world and the hereafter. In other words, people need to act in accordance with Islamic law, which has been outlined in the Qur'an and Sunnah. Therefore, an ideal society meets the needs of humans living on and as well as the hereafter by subscribing to the laws of Allah.

3. Discussion

The concept of an "ideal society" is analyzed through the prism of the so-called "philosophy of Illumination". The concept of "linguistic identity" allows one to study a person as a representative of society and as a unique self-valuable person. In such a paradigm, the person is recognized as a measure of all things.

The author is focused on the process of creating a holistic concept of the relation between tradition and thinking. In the framework of the given research, new interpretations of the elements of political discourse are presented. The "semantic power" that Imam Khomeini reached earlier was actually a basic tool for the development of society. People are integral to the formation and continuity of semantic structures of power. Within this framework, the concept of power is examined as a form of legal institutional discourse. This analysis suggests that Khomeini's later political theory, *Wilayat al-Faqih* can be legitimized on the basis of gnostic and metaphysical principles, an important consideration for future research.

In today's world, characterized by the globalization of communication and emotions, appealing to hearts and minds can serve as a particularly influential tool. This study examines the relationship between the philosophy of Illumination and the Islamic concept of community through a philosophical lens, an approach that contributes to the novelty of the research. Philosophy, as the primary method of inquiry, is particularly attentive to internal distinctions and nuances, which provide an important explanatory framework for understanding religious changes and their social, cultural, and political implications.

4. Conclusion

Khomeini developed a contextual approach to interpreting the Qur'an and its socio-legal passages based on their specific social and historical contexts. Imam Khomeini brought forward innovative ideas in realizing the concept of an "ideal society" by putting religion at a high level. In order to make religion the foundation of national development, the semantic power was applied. The state has to streamline its system through application of religious laws. Good and fair leadership is vital to achieving harmony. The "ideal society" is one based on fairness and dignified living where an important part of the Quran's purpose is revealed. He develops the idea for an "ideal society" as a new type of collaborative society based on knowledge and values, where all goods are distributed between citizens due to existing principal of solidarity.

References

Aboul-Enein, B. H. (2016). Health-promoting verses as mentioned in the Holy Quran. *Journal of Religion and Health*. 55(3), 821–829. https://doi.org/10.1007/s10943-014-9857-8.

Akhmetova, E. (2017). Al-Farabi and Said Nursi on the Civilizing Mission of the Prophets. *Intellectual Discourse*, 25(Special Is): 453–475.

Arzhanovskaya, A. E. Eltanskaya, Y. Linkova, L. Medvedeva. (2018). Representation of Semantic Power in Discourse of Institutionality. *SHS, Web of Conferences 50*, 01030. https://doi.org/10.1051/shsconf/20185001030.

Bazzi. E. (2023). The Mystic as a Political Leader: The Impact of 'Irfān' on Imam Khomeini's Political Doctrine. *Journal of Contemporary Study of Islam*. 3(2), 117–125. https://doi.org/10.37264/jcsi.v3i2.01.

Chittick, W. (1989). *The Sufi Path of Knowledge: Ibn al-Arabi's Metaphysics of Imagination*. Albany, NY: SUNY.

Corbin, H. (1971). *En Islam Iranien*. V. 1. Aspects spirituels et philosophiques. Paris: Gallimard.

Damad, M. (2000). *The Spiritual Sovereignty of the Perfect Man, Message of Thaqalayn* 5 (3): 34–38.

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*. 62(1): 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Eltanskaya, O. Popova, Yu. Linkova, A. Arzhanovskaya, Yu. Kulichenko. (2018). ASSEHR 97: 100–102.

Eltanskaya, E. (2011). *The discourse medium of spatial dynamics*. Volgograd scientific publishing house.

al-Farabı, Abu Nasr. (1985). *Perfect state: Kitab mabadi' a ra' ahl al-madına al-fadila*. On the perfect state, a revised text with introduction, translation, and commentary, by Walzer R. Clarendon Press, Oxford.

Fehri, A. (introd.). (1982). *Ayatollah Khomeini, Sharh Dua'a al-Sahar*. Beirut: Muassat al-Wafaa.

Gutas, D. (2004). *The meaning of madanı in al-Farabı's 'political' philosophy*. Mélanges de l'Université Saint-Joseph 57: 269–281.

Hummel, C. (1994). Plato. *Prospects*, 24(1–2): 329–342. https://doi.org/10.1007/BF02199024.

Kamal, Z. (1995). Demensi akhlak dalam masyarakat utama [moral dimensions in the main society]. In M. Yunus Yusuf (ed.) *Masyarakat utama: konsep dan strategi* [the main society: concepts and strategies]. Jakarta: Perkasa.

Khamenei, S. M. (2000). *Development of Wisdom in Iran and in the World*, Sadra Islamic Philosophy Research Institute.

Khomeini R. (2007). *Misbah al-Hedaya ila al-Khilafah wa al-Wilayah* [the Lamp of Guidance into Viceregency and Sanctity]. The Institute for Compilation and Publication of Imam Khomeini's Works (International Affairs Department), Tehran.

Khomeini, R. (1990) *Sahife-ye Nur* [Anthological Compendium] [Sazman Asnad wa Madarek Enghlab-e Eslami, V. 18.

Khomeini, E. (1986). *Taaliqat ala Sharh Fusus al-Hikam wa Misbah al-Uns* [Glosses on the Explication of the Bezels of Wisdom and the Lamp of Proximity], Tehran.

Khomeini, R. (1983). *Misbah al-Hedaya ila al-khilafah wa al-Wilayah* [al-Wafaa, edition].

Lubna, A. (2008). Post-Revolutionary Iran. *Foreign Policy*, University of Punjab, Lahore.

Mutahhari, M. (1986). *Masyarakat dan Sejarah* (Society and History). Jakarta: Mizan.

Mutahhari, M. (1983). *Insan-i Kamil* (The Perfect Man), Qum: Intisharat-i Islami.

Nasr, H. (1972). Shiism and Sufism: Their Relationship in Essence and in History. (ed.) Hosein Nasr, *Sufi Essays*. Albany, NY: SUNY Press.

Nasr, V. (2006). When the Shiites Rise. *Foreign Affairs*. New York. July/August.

Netton, I. (1985). Al-Farabi on the Perfect State: Abū Naṣr Al-Fārābī's Mabādi' Ārā Ahl al-Madīna al-Fāḍila. A revised text with introduction, translation and commentary by Richard Walzer, Oxford, Clarendon Press, 1985. *Journal of the Royal Asiatic Society of Great Britain & Ireland*, 118(1).

Plato. (2021). *The Republic*. Chartwell Books.

Rahman, F. (1987). ASFĀR AL-ARBA'A. (The four journeys), title of the magnum opus of Mollā Ṣadrā. Vol. II, Fasc. 7, 744–747.

Ridgeon, L. (2014). Hidden Khomeini: Mysticism and Poetry. Arshin Adib-Moghaddam (ed.), *A Critical Introduction to Khomeini*. Cambridge: Cambridge University Press: 193–203.

Sachedina, A. (1998). *The Just Ruler in Shiite Islam*. Oxford: Oxford University Press.

Ṣadr al-Dīn Shīrāzī. M. (1964). *Kitab-i al-Masha'ir*. Qismat-i Iranshinasi Institu-i Iran va Faransah, v.10.Bibliothèque Iranienne.

Schreier, M. (2012). *Qualitative content analysis in practice*. Thousand Oaks, CA: Sage.

Sheigal, E. (2004). Power as a concept and category of a discourse. Mode of access: www.gumer.info.pdf, JournalofLegaleducation 34 (3).

Sycheva, C. G. (2000). *The Problem of the Symbol in Philosophy*, Tomsk University Press.

Usman, A. H., & Kadir, M. N. A. (2020). Discussion on the Maqaṣid al-Qur'an: Classical and Contemporary Intellectual Response. al-Irsyad. *Journal of Islamic and Contemporary Issues*, 5(1), 188–196. http://al-irsyad.kuis.edu.my/index.php/alirsyad/article/view/67.

Walzer, R. (1967). *Al-Farabı and his successors. The Cambridge history of later Greek and early medieval philosophy*, (ed.) Armstrong AH. Cambridge University Press, Cambridge, 652–669.

Whitehead, A. N. (1999). *Symbolism, its meaning and impact*, Tomsk University Press.

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 283 – 292

PERSONHOOD IN AFRICAN ENVIRONMENTAL ETHICS

Terzungwe Inja*

Rotimi Omosulu**

Abstract: African environmental ethics has come a long way in advancing theories that would, (i) prove the existence of African philosophy to the rest of the world; (ii) show the efficacy of age-old African practices in sustaining the environment; and (iii) point out the African non-anthropocentric path towards preservation, conservation, and sustainability of the environment. However, there seems to be a misunderstanding of personhood as it applies to African environmental ethics, leading to the claim that the latter is essentially anthropocentric and offers no unique contribution to environmental ethics in general. Thus, a sound understanding of personhood in African thought would help to clarify the kinds of value, rights, and worth that are extendable to nature within African environmental ethics. It is against this backdrop that this paper highlights how personhood serves as the ontological foundation of African environmental ethics. To achieve this goal, the paper examines the concept and status of personhood within the ambit of three theories of African environmental ethics: (i) Segun Ogungbemi's Ethics of Nature-Relatedness, (ii) Godfrey Tangwa's Eco-bio-communitarianism, and (iii) Mogobe Bernard Ramose's Ubuntu Ecology. The paper concludes that personhood lies at the root of the examined theories of African environmental ethics and determines the kind of worth they have extended to nature.

Keywords: personhood, ethics of nature-relatedness, eco-bio-communitarianism, Ubuntu Ecology, African environmental ethics.

1. Introduction

The nature of environmental challenges in Africa is in several ways different from that of other parts of the world. This explains why J. Baird Callicott (1994) argues that African environmental practices are in many ways different from those of the West.[1] Workineh Kelbessa (n.d.) also notes that due to colonialism and other factors in Africa, traditional philosophies which informed the environmental practices of the people, are mostly unrecognised and not considered by policy makers in Africa. What follows is the failure of Western-based environmental ethics and management policies in Africa (p. 31). Meanwhile, scholars in many disciplines have argued in support of the study of African ancestral wisdom, worldview, and traditional practices as a solution to Africa's peculiar problems. For instance, in the field of African social

[1]See for instance, Baird J Callicott, *Earth's Insights: A Multicultural Survey of Ecological Ethics from the Mediterranean Basin to the Australian Outback* (Berkeley: University of California Press, 1994).

*Department of Philosophy, Benue State University, Makurdi **Department of Language, Linguistics and Philosophy, University of the West Indies, Mona Campus, Kingston, Jamaica; rotimi.omosulu@uwi.edu

©2026 Inja and Omusulu. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

and political philosophy, the issue of personhood and how it determines a meaningful life for the African has been applied (by some scholars) to buttress its connection to the socio-political life of the African.[2] Similarly, Bruno Yammeluan Ikuli, and Peter Lambert Ukanag (2021) have also examined personhood in relation to human rights within the African communitarian setting (p. 211).

This paper propels the conversation further by looking at how the African notion of personhood forms the foundation of some African theories of environmental ethics in today's society. It is an explorative journey into unravelling the nature or 'rights and privileges' which humans extend to nature in the African version of non-anthropocentric ethics. Hence, the paper contains a conceptualisation of personhood in African thought, and the notion of African environmental ethics. It further examines three theories of African environmental ethics in order to unearth their conception of personhood and how it applies to their extension of ethical worth to nature.

2. Understanding personhood in African thought

This section discusses personhood in African traditional thought without grounding the discussion in the Western philosophical paradigm of personhood as some scholars would have expected. In that way, what a person is in African traditional thought would be understood genuinely from the African perspective. It is with this in mind that we say there is no unanimity in African thought on the particularities of what personhood entails – not in terminology, nomenclature, ontology, or expression. Nevertheless, it is possible to outline certain common features as one pans from one African culture to the other. Scholars like Bolaji Idowu, John S. Mbiti, Kwasi Wiredu,[3] Kwame Gyeke, Kelvin Behrens, among several others, have already exposed the diversity of African thought on personhood from religious, sociological, anthropological, and philosophical perspectives (Molefe, 2020a, p. 197). Personhood has also been viewed from descriptive and normative, ontological, social, and existential perspectives (Ikuenobe, 2015, p. 1006; Molefe, 2020a, p. 3). Factors such as reasoning, consciousness, moral conduct, engagement with one's community, property rights, social recognition, legal and moral obligations/responsibilities, are seen as constituents of personhood in Africa. For example, Molefe (2020b) draws upon the ethical dimension of personhood in Africa by saying that:

> [...] the ethics of personhood is constituted by three components, namely (1) the fact of being human, which informs (2) a view of moral status qua the capacity for moral virtue, and (3) which specifies the final good of achieving or developing a morally virtuous character...[In the light of this] a meaningful life is a function of achieving moral excellence or perfection. The moral perfection proposed here [...] is of a deontological and satisficing kind. [This shows that] The achievement of satisfactory levels of moral excellence, within sociopolitical and moral limits, captures the essence of a meaningful life (p. 195).

[2]See for instance Bernard Matolino, "The Concept of Person in African Political Philosophy: An Analytical and Evaluative Study" (University of KwaZulu Natal, Pietermarizburg Campus, 2008).

[3]See "Personhood in African Thought," New Dictionary of the History of Ideas, 2024, n.p., https://www.encyclopedia.com.

From the above, we argue that personhood in Africa revolves around a communitarian worldview which stands on three legs like a tripod: the socio-political, ethical, and psycho-spiritual. A person is the community, just like the community is the person; and yet there is no contradiction in the roles that both parties play. In other words, personhood in Africa is a matter of relations, as against a relationship built on self-centredness, exploitation and moral apathy. This means that African personhood is not constructed to promote a worldview that harbours survival of the fittest. Both in theory and practice, African personhood gears only towards fulfilling the interests of the community and its members, even though the interest of the community often overrides that of a person whenever there is a clash of interests. If, for instance, a man is interested in marrying a lady from a community that is well-known for a particular health issue, such as smallpox, his community could prevail on him to relinquish such an interest in order to save the community from such a communicable disease.[4] In this light, safeguarding a community's interest against a personal interest should not be understood here as the tyranny of the majority in African personhood, for the community and its members are thoroughly educated (i.e., indigenous education) on the benefits that follow the prioritisation of communal interest over personal interest. Social interdependence for the benefit of all members of a community is the end that African personhood seeks vigorously.

In this connection, individual autonomy is neither lost nor are the communal rights violated in African personhood (Ikuenobe, 2015, p. 1005). Rather, personhood in Africa presents a complementary interaction between a community and a person, just like that between two hands washing each other in a bowl of water. If a tree does not make a forest, then a person does not make a community in African personhood. In the case of the symbiotic relationship that exists between a person and a community in ancient Africa, each member of an African community has the moral obligation to put the value of 'we' above the value of 'I', especially as regards a choice of actions and inactions. As Rotimi Omosulu (2022) argues, the African value of 'we' "encapsulates unflinching kinship solidarity by a people bound together by common destiny…what drives communal ethics in indigenous African societies is the value of 'we'" (p. 8). Given that African personhood thrives on the value of 'we', which is also expressed in Ubuntu ethics ('I am because we are'), it would amount to a display of ignorance should anyone conclude that personhood in Africa fails to recognise individual autonomy. On the contrary, a person is a person in the ancient African society if she chooses actions that encourage social harmony and communal flourishing over actions that promote individual wellbeing at the detriment of collective wellbeing. Therefore, personhood in Africa gives room for moral accountability because the interdependence of life does not suggest the suspension of moral responsibility by a moral agent in an ancient African society. Let us now consider the implication of this notion of personhood (in African thought) on African environmental ethics.

[4]The authors are aware that Smallpox, as a global health disease, was eradicated by the World Health Organization in 1980. It is only referenced in this study to drive home how the interest of the community supersedes that of an individual.

3. Personhood in African environmental ethics

It is no longer in dispute that African philosophy differs from that of other regions of the world due to the cultural, social, economic, and geopolitical experiences of the African (Chimakonam, 2024, p. n.p). Thus, African environmental ethics may be defined as the ethical inquiry into Africans' particular ways of conceiving, perceiving, contemplating, and interacting with the environment; as well as the moral status they have accorded to nature. Of course, earliest discourse on African environmental ethics has "been determined by the direction of African Philosophy…especially after the second half of the twentieth century" when the opinion was rife that Africans had no elaborate ethics, except for various myths and religio-cultural rituals (Chemhuru, 2019, p. 1). However, the twenty-first century has seen a thriving of African environmental ethics in quite a broad range of variety. Despite the progress, scholars like J. Baird Callicott (Callicott, 1994) still hold a negative view against African environmental ethics by dismissing it as essentially anthropocentric. Thus, we shall explore the notion of personhood in some theories of African environmental ethics in order to dispel this misconception.

3.1. Ethics of Nature-Relatedness

In his "An African Perspective on the Environmental Crisis", the Nigeria-born Segun Ogungbemi proposed a theory called the Ethics of Nature-Relatedness, as a viable solution to the contemporary environmental degradation in Africa. Having examined the communitarian nature of African societies, Ogungbemi (1997) notes that in modern Africa, exploitation of land (and other natural resources) was contrary to the traditional African worldview on the environment. According to him:

> In our traditional relationship with nature, men and women recognize the importance of water and air management to our traditional communities. The ethics of not taking more than you need from nature is a moral code. Perhaps this explains why earth, forest, rivers, wind and other national objects are traditionally believed to be both natural and divine. The philosophy behind this belief may not necessarily be religious but a natural means by which the human environment can be preserved […] the ethics of care is essential to [decode the] traditional [African's] understanding of environmental protection and conservation (p. 206).

He goes further to argue that African environmental thought contains principles that are universally applicable to environmental management in contemporary times. So, he draws on these principles to coin what he calls the 'ethics of nature-relatedness'. This, he explains, is an ethics of care rooted in traditional African ways of interacting with the environment as mentioned in the above quotation. The theory acknowledges that nature does not need man to exist and function. It is an ethic that obliges man to approach nature with reasonable concern for its worth, survival, and sustainability (pp. 208–9). This ethic, to Ogungbemi, comprises three key elements: reason, experience, and the will; and it does not attribute spirituality to natural resources. In practice,

Ogungbemi concludes that nature-relatedness could help in curbing the environmental crisis of today in a manner in which the Western paradigms may fall short (p. 209). Ogungbemi's theory has been criticised on various grounds, including that his thought is rather anthropocentric, is not sufficiently African, and amounts to "an alienation of the African spirit and peculiar experiences…[as well as a] disconnectedness…from the African ontology" (Edor, Idagu, & Ogelenya, 2025, p. 13). But these are not our concerns; the focus here is on how his theory showcases the role of personhood in understanding African environmental ethics.

Even though Ogungbemi's work has been critically engaged by Philomena Aku Ojomo (2011), we argue that his theory of 'nature-relatedness' outlines the nexus between African personhood and African environmental ethics. This is so because the communitarian nature of African personhood places an obligation on humans to treat nature as part of the community; and as such, moral worth is extended to nature in ancient African communities since the vitality of a community is tied to the vitality of nature. In other words, there is a necessary ontological relationship between humans and nature in African personhood, especially if the deontological content of African personhood is well comprehended. It is this 'relatedness' that Ogungbemi expresses as an 'ethics of care'. The African extends the care he has for other persons in the community to nature because he does conceive of nature as a 'person' worthy of such care – not as a thing to be exploited. From this, it is inferable that Ogungbemi's Ethics of Nature-Relatedness emphasises the interwovenness between personhood and environmental ethics in the African worldview.

3.2. Eco-bio-communitarianism

The Cameroonian, Godfrey Tangwa, in his 2004 article, "Some African Reflections on Biomedical and Environmental Ethics", presented Eco-bio-communitarianism in a way that made him a voice to be reckoned with in the 21st century conversation on African environmental ethics. One of the contents of his theory of Eco-bio-communitarianism is the attitude of "live and let live". He argues that both anthropocentric and non-anthropocentric ethics could endanger nature if they fail to adopt a "live and let live" attitude. However, Tangwa reveals why the African environmental worldview could not be dangerous to nature in general. This is because the traditional Africans:

> tend to be more cosmically humble and therefore not only more respectful of other people but also more cautious in their attitude to plants, animals, and inanimate things, and to the various invisible forces of the world. One might say, in short, that they are more disposed toward an attitude of live and let live (p. 389).

He goes further to showcase how the pre-colonial Africans' attitude to nature is rooted in their (Africans') metaphysical worldview. According to him, the African metaphysical world emphasises the "recognition and acceptance of interdependence and peaceful coexistence between earth, plants, animals and humans"; in contrast

with the anthropocentric and individualistic view of the West (Tangwa, 2004, p. 389). Tangwa holds that in the African metaphysical worldview, the dichotomy between "plants, animals, and inanimate things, between the sacred and the profane, matter and spirit, the communal and the individual is a slim and flexible one" (p. 389). He explains that it is only within the context of this metaphysical outlook that one could comprehend the people's belief in the transmigration of the human soul into animals, or plants, or into forces such as the wind.

Based on the foregoing, Tangwa disapproves of Western environmental attitudes which are mostly backed by aggressive and dominating acts towards nature. Also, he interprets the elaborate rituals of by which Africans frequently commune with God, divine spirits, departed ancestors, and sundry forces of nature as indicative of an attitude of respectful co-existence, conciliation, and containment, rather than a domineering and exploitative one (p. 390). From this point, Tangwa discards not only modern technology itself, but also the motivation for its development. He objects to the idea that grants the use of Western technology mainly to satisfy mankind's desire to dominate the world. Eco-bio-communitarianism, for him, is the plausible means to achieve environmental sustainability in this contemporary time (p. 394).

Responding to Tangwa's position, Felix Ayemere Airoboman (2019) identifies two problems. The first is Tangwa's "making of copious distinctions between man and nature"; and the second being the "identification of asymmetrical moral relation between man and the rest of nature as the justification for human use of nature" (p. 109). One may only but agree that these are serious issues considering the fact that both problems seem to be angling towards anthropocentrism, thereby threatening the very non-anthropocentric foundation (of African environmental ethics) which Tangwa has set out to defend. Without delving into that debate, it suffices to state here that, in spite of those so-called flaws, Tangwa's eco-bio-communitarianism showcases the communitarian nature of personhood in African thought – by which one is only a person if he seeks to 'live and let live' in service and harmony with all facets of the community. His depiction of the African's interaction with nature clearly demonstrates the kind of 'personhood' which African conceives of, in relation to their environment. The metaphysical attitude of care that exists between humans and nature depicts a communitarian and 'transcendent' personhood by which the sharp boundaries between mankind and other beings are blurred out. Thus, the person is viewed as one who interacts with all spheres of the cosmos bearing an attitude of respect, harmony, and interdependence. This too, highlights a clear extension of moral worth from mankind to non-human beings.

3.3. Ubuntu Ecology

Mogobe Bernard Ramose is a renowned South African philosopher. He situates his environmental theory on the Bantu philosophy of Ubuntu. Ubuntu, which means 'humanity' evolves from the Bantu idea of 'I am because we are'. In his article titled, "Ecology through Ubuntu", Ramose (2015) differentiates between Ubuntu and Western humanism. To him, the term humanness better explains Ubuntu because it alludes

to "both a condition of being and the state of becoming…", unlike humanism which suggests "a condition of finality, a closedness or a kind of absolute…" (pp. 69 – 76). Further, he argues that this ontological difference (between the African humanness and Western humanism) is at the root of the divergence in attitudes towards physical nature between the two philosophical traditions.

Ubuntu as humanness, Ramose explains, entails the attitude of seeing the universe "as the unceasing unfoldment of interaction and interdependence between and among all that there i" (p. 74). In this context, "all that there is" embodies an all-encompassing view of reality where nature and humans converged to complement each other because none could exist without the other. Thus, Ramose affirms that Western technology is the root cause of the contemporary ecological crisis. According to him, "technological advancement continues to reaffirm the need to restore botho because more than ever before, humanity is faced with the threat of catastrophic ecological disaster. This is exemplified by the widespread air pollution, global warming, the destruction of the ozone layer and the ever constant [sic] threat of nuclear omnicide" (p. 75).

Using the Sotho proverb, *motho ke motho ka batho* ('to be human, recognise the humanity in others, and so be humane towards them') Ramose argues that *botho* (being humane to others) constitutes the core of extending ubuntu towards the environment. Just as both the individual and the community need each other to thrive, so also do humans need the environment to survive. Similarly, just as humans owe one another a duty of care, humans also owe an obligation of care to physical nature (p. 76). In other words, Ramose sees humans as part of nature and nature as part of humans due to the relationship of interdependence that exists between them. Having identified Western Technology as the main culprit of ecological destruction, Ramose also condemns the separation of *botho* (humanness) from the invention and use of Western technology. To turn the tide, he recommends the restoration of *botho* (p. 76). Ubuntu/botho is for Ramose, the appropriate philosophy that could be used to construct a non-anthropocentric relationship between humans and nature in this contemporary time.

Meanwhile, what is present in Ramose's Ubuntu-inspired environmental philosophy is the notion of African personhood. This is so because the African ontological and deontological conceptions of a person could be easily deciphered in Ramose's argument for environmental sustainability. According to Samuel and Omosulu (2024), personhood in Ubuntu could be understood from ontological and ethical senses. The former sense says that "a person is a human being who is who she becomes partly by virtue of her relationship with other people and beings", while the latter sense affirms that "a person is someone who acts with appropriate moral value…" (p. 159). Therefore, Ramose's dependence on *botho* to advocate for environmental sustainability suggests two things: (1) that relationship is a pivotal aspect of human life and should be respected; (2) that an inappropriate relationship either the one among human beings or the one between human and nature should be corrected through Ubuntu education. This means that Ramose is not only using the notion of personhood in Ubuntu to promote peaceful coexistence among humans, but he is also relying on it

to construct a relationship of respect between humans and nature. In other words, Ramose's reliance on *botho* as a remedy to the environmental crisis in Africa and beyond resonates with a notion of personhood that deliberately extends moral worth to nature without reservations.

4. Personhood and Nature's worth

At this juncture, a careful look at the views of African environmental scholars discussed above shows the one thread that runs through all of them; that is, the peculiar form of personhood in Africa – which is rather communitarian in scope. They all directly or indirectly outline personhood as interaction, interdependency, communitarianism and most importantly, an extension of moral worth to non-human aspects of the cosmos.

This kind of personhood dictates an ethical duty which is not a separate idea but a natural and necessary extension of the obligation to care for one's community. In turn, the community in question includes the living, the dead, the future generations, spirit forces, and of course, the entire web of life. Without conceding the sharp disparities in their socio-political origins, one may state that the African views of nature are in some ways related to Western biocentrism and ecocentrism. While biocentrists grant the extension of moral worth to all organisms based on their sacredness thereof, ecocentrists support the extension of moral worth to the entire ecosystem. However, both schools of thought agree that man is only but a part of the ecosystem and not a superior, all-important controller (Amendolare, 2024).

Furthermore, it must be recalled that personhood in Africa is not homogeneous, but it is the binding force within the environmental ethics of African communities. It determines the rights, privileges, and moral worth extended to nature by the African. Thus, when the African approaches physical nature, she does so as though she were encountering a person to whom she owes a duty of care. It is not just a care for the other, but a care that includes oneself. The understanding is that to hurt another is to hurt oneself. Thus, Mbih Jerome Tosam (2019) reports that among the Kom of Cameroun, it is forbidden to cut any plant at night using a machete. But if one must do so, she must make the incantation, "*yi leu'sɨ itu ɨyvɨ*", translated as 'take away your heads.' This statement, Tosam explains, "is aimed at spirits who are believed to incarnate plants at night and if wounded may be reborn with cleft lips" (p. 178).

So, through this peculiar conception of personhood, the African understands, not just her interdependence with nature, but her interconnectivity. Thus, she treats nature as she would treat herself. The African, through taboos and restrictions, shows that she must not take more from nature than she needs, must not involve in wanton destruction, and definitely must not seek to dominate nature just for the sake of self-glorification or the fun of it. In this light, the moral worth extended towards nature by the African ethic is definitely not anthropocentric as alleged by those who misconstrue the African nature of personhood.

5. Conclusion

The above exercise was concerned with ascertaining the role of personhood as a factor that underlies African environmental ethics. As it turned out, the African notion of personhood was found to be rooted in communitarianism, which is a central factor in the extension of moral worth to nature. This realization, it is hoped, should put to rest the notion that African theories of environmental ethics are merely anthropocentric.

References

Airoboman, F. A. (2019). Metaphysical Foundations of Environmental Ethics in an African Culture. *NAJOP: Nasara Journal of Philosophy, 4*(1), 96–117.

Amendolare, N. (2024). Ecocentric vs. Biocentric Philosophies. Retrieved May 8, 2024, from: https://study.com/academy/lesson/differences-between-ecocentric-biocentric-philosophies.html

Callicott, B. J. (1994). *Earth's Insights: A Multicultural Survey of Ecological Ethics from the Mediterranean Basin to the Australian Outback*. Berkeley: University of California Press.

Chemhuru, M. (2019). Introducing African Environmental Ethics. In M. Chemhuru (Ed.), *African Environmental Ethics: A Critical Reader* (pp. 1–5).

Chimakonam, J. O. (2024). History of African Philosophy. In *Internet Encyclopedia of Philosophy* (p. n.p.). Retrieved from https://iep.utm.edu/history-of-african-philosophy/

Edor, E. J., Idagu, U. A., & Ogelenya, G. (2025). A Critical Analysis of Ogungbemi's Ethics of Nature-Relatedness and Tangwa's Eco-Bio-Communitarianism. *Studia Ecologiae et Bioethicae, 23*. https://doi.org/http://doi.org/10.21697/seb.5849

Ikuenobe, P. (2015). Relational Autonomy, Personhood, and African Traditions. *Philosophy East and West, 65*(4), 1005–1029. Retrieved from https://www.jstor.org/stable/43831225

Ikuli, Y. B., & Ukanag, P. L. (2021). A Study of Kwame Gyeke's Notion of Personhood and Its Relationship to Rights in African Communitarian Structure. *International Journal of New Economics and Social Sciences, 2*(14), 211–229. Retrieved from https://ijoness.com/

Kelbessa, W. (n.d.). Can African Environmental Ethics Contribute to Environmental Policy in Africa? *Environmental Ethics, 36*(1), 31–61. https://doi.org/https://doi.org/10.5840/enviroethics20143614

Molefe, M. (2020a). *African Personhood and Applied Ethics*. Makhanda, South Africa: African Humanities Program.

Molefe, M. (2020b). Personhood and a Meaningful Life in African Philosophy. *South African Journal of Philosophy, 39*(2), 194–207. https://doi.org/https://doi.org/10.1080/02580136.2020.1774980

Ogungbemi, S. (1997). An African Perspective on the Environmental Crisis. In L. J. Pojman (Ed.), *Environmental Ethics: Readings in Theory and Application*. Belmont, CA: Wadsworth Publishing Company.

Ojomo, P. A. (2011). Environmental Ethics: An African Understanding, *Journal of Pan-African Studies*, 4, 101–113.

Omosulu, R. (2022). Western and African Perspectives on Value. *NAJOP: Nasara Journal of Philosophy, 7*(2), 1–19.

Ramose, B. M. (2015). Ecology through Ubuntu. In M. Roman (Ed.), *Environmental Values: Emerging from Cultures and Religions of the ASEAN Region* (pp. 69–76). Bangkok: Guna Chakra Research Centre, Department of Philosophy and Religion, Assumption University.

Samuel, O. S., & Omosulu, R. (2024). Ubuntu Thinking on Biodiversity Loss: The Inadequacies of Communitarian and Egalitarian Solutions. *Journal of Applied Philosophy, 41*(1), 145–169.

Tangwa, G. B. (2004). Some African Reflections on Biomedical and Environmental Ethics. In W. Kwasi (Ed.), *A Companion to African Philosophy*. Oxford: Blackwell Publishers.

Tosam, M. J. (2019). African Environmental Ethics and Sustainable Development. *Open Journal of Philosophy, 9*, 172–192. Retrieved from http://www.scirp.org/journal/ojpp

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 293 – 311

MAIN FEATURES OF EXISTENTIALISM IN FRENCH LITERATURE (BASED ON THE WORKS OF J.M.G. LE CLÉZIO AND P. MODIANO)

Rasima Asadzade*

Abstract: The study examines existential motifs in the works of Le Clézio and Modiano and to analyse their interpretation within French literature of the 20th and 21st centuries. The research employed comparative, analytical and structural methods to examine the artistic representation of existential ideas, focusing on key motifs, narrative strategies and semantic interpretations of loneliness, alienation, identity crisis and memory. The findings show that existential motifs develop in different directions in the works of the two authors. In Le Clézio's texts, loneliness and alienation are linked to the loss of cultural roots, the opposition between natural and urban spaces, and the conflict between tradition and civilisation. His characters experience identity crises expressed through a search for authenticity and belonging. In contrast, Modiano's works associate existential crisis with fragmented memory, the impossibility of reconstructing a coherent biography, and social alienation. His narrative structures emphasise uncertainty, while the motif of oblivion intensifies a sense of existential instability. The comparative analysis confirms that both writers address loneliness and alienation as fundamental aspects of human existence, yet their interpretations are shaped by distinct social and historical contexts. Le Clézio focuses on space and cultural identity, whereas Modiano highlights memory and the erosion of the past.

Keywords: identity crisis, alienation, loss, desire for authenticity, fragmentation.

1. Introduction

Existentialism was one of the most substantial philosophical and literary movements of the 20th century, having a significant impact on the development of Western culture. Its key concepts, such as absurdity, freedom, responsibility, loneliness and identity crisis, were reflected in the works of many writers. French literature has become one of the main platforms for the comprehension of existential ideas. In the first half of the 20th century, existential themes were articulated by Sartre (1905-1980), Camus (1913-1960), and Beckett (1906-1989), whereas in the second half of the 20th century and into the 21st century, they were reinterpreted and expanded in the works of Le Clézio (b. 1940) and Modiano (b. 1945). Their works include the themes of alienation, memory, personal crisis and the search for meaning, which makes their works valuable material for analysing the existential aspects of contemporary French literature.

The research relevance is determined by the need for a deeper analysis of the manifestations of existentialism in the works of Le Clézio and Modiano. Despite the considerable attention paid to existential philosophy and literature, specific aspects of

*Department of World Literature and Theory, Azerbaijan University of Languages, Azerbaijan; rasimaasadzade179@gmail.com

©2026 Rasima Asadzade. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

existential motifs in their works remain insufficiently explored. In the contemporary cultural context, transformations in the perception of identity, memory, and existence have significantly influenced literary representations of human experience. Recent studies indicate that modern literary texts increasingly portray identity as unstable and shaped by fragmented cultural and historical memory rather than by coherent personal narratives (Phillips, 2020; Wertz, 2023). Consequently, existential motifs of alienation, uncertainty, and the search for authenticity are articulated not only at the level of individual experience but also in relation to broader socio-cultural processes. As Bondar (2022) notes, contemporary literature tends to reinterpret classical existential concepts in response to changing historical and cultural conditions. In this context, the analysis of the works of Le Clézio and Modiano contributes to a more precise understanding of the transformation of existential ideas in contemporary French literature. These transformations reflect a shift from stable identity models to fluid and fragmented forms of subjectivity in late modernity.

Cropper and Browne (2024) analysed nihilism, anxiety and identity crisis in late 20th-century French philosophy. The study identified the relationship between existentialism and the social and cultural processes of this period and concluded that existentialist issues continued to influence literature, but in a modified form. Gosetti-Ferencei (2020) examined existentialism in literature through the prism of human identity formation. The study analysed the works of several writers, emphasising their philosophical significance. The study identified key literary techniques that contribute to the transmission of existential ideas. Mahajan (2021) studied the philosophical aspect of existentialism in literature, considering the basic principles of this trend and their impact on narrative structures. The study concluded that existential literature offered a special way of analysing human existence.

Rahimipour (2021) conducted a comparative analysis of the concept of existence in the works of Sartre and Hemingway, prioritising the educational aspects of existentialism. The study concluded that existential philosophy influenced not only fiction but also the formation of a person's worldview. Girard (2024) analysed the influence of existentialist philosophy on French society. The study argued that individualism and existential motives continued to be significant in the culture of the 21st century. Quindadiarto (2023) studied the identity crisis and alienation in literature in the context of existentialism, highlighting *The Metamorphosis* by Kafka. The study concluded that the concepts of personality and existence evolved in the literature of the 20th century. Guignon (2020) studied the moral aspects of existentialism, analysing the ethical concepts of this trend. The study emphasised that existential ethics remained relevant and continued to develop. Sasan (2023) examined the influence of existentialism on the research of knowledge, morality, values, and religion. The study argued that existential ideas continued to shape views on human existence.

McElroy (2022) conceptualised existential criticism as an approach that examines the representation of freedom, existential choice, alienation, authenticity, and the experience of absurdity in literary texts, emphasising the role of narrative structures, character consciousness, and thematic patterns in articulating existential meanings.

Zhou (2024) explored the relationship between existentialism and authenticity in contemporary social discourse, highlighting the relevance of existential concepts for understanding individual experience, although literary texts were not the focus of the analysis.

In this study, existential motifs are identified through the analysis of narrative structure, character development, and thematic organisation. At the structural level, attention is paid to narrative fragmentation and temporal discontinuity; at the level of characters, to moments of existential choice, identity crisis, and alienation; and at the thematic level, to recurring motifs of loneliness, memory, and the search for authenticity. Narrative techniques such as spatial symbolism and autobiographical elements are examined as mechanisms through which existential meanings are articulated in the texts.

The analysis of existing studies has revealed several gaps. Most studies are devoted either to the philosophical aspects of existentialism or to the analysis of classical existential literature. Studies aimed to analyse existential motifs in contemporary French literature remain limited, and the works of Le Clézio and Modiano are not sufficiently covered in this context. Thus, there is a need to study existential motifs in their works, which would not only identify the peculiarities of their artistic comprehension of existentialism, but also determine the place of their work in the development of French literature of the 20th and 21st centuries.

Thus, this study aims to identify the characteristics of existential motifs in the works of Le Clézio and Modiano, to examine their connection with classical existential concepts, and to analyse the transformation of these ideas in contemporary literature.

2. Materials and methods

The study prioritised the identification of existential concepts in philosophical works and their artistic embodiment in the works of Le Clézio and Modiano, which established links between the philosophy of existentialism and contemporary literary texts.

The analysis of philosophical works was a key stage of the research, as it was impossible to define the conceptual framework of the existentialist tradition without it. The study included works by Heidegger (2010); Jaspers (1971); Sartre (2007); and Camus (2013). These sources were selected because of their significance for the formation of existentialist thought. Heidegger's work was used to examine the concept of Dasein and the problem of authentic existence, which provided a theoretical basis for analysing the crisis of identity and human abandonment in literary texts. Jaspers' research was used to study situations and transcendence as ways of analysing existence, which influenced the interpretation of the characters in the works of Le Clézio and Modiano. Sartre's work served as a source for analysing the concepts of freedom, responsibility and "bad faith", which were reflected in the literary characters. Camus' work was included in the study to examine the concepts of absurdity and existential rebellion, which are substantial in the study of existential motifs in contemporary literature.

The analysis of the works was based on the textual method, which identified existential motifs in their narrative structure, character system and key themes. The choice of Le Clézio and Modiano was due to the fact that they represent the most striking examples of the artistic transformation of existential philosophy in the literature of the late 20th and 21st centuries. The study included the novels by Le Clézio: *L'Africain* (2010), *Desert* (2010) and *La Quarantaine* (1995). These works were chosen due to their thematic focus on the problem of identity, the search for roots and reflections on the place of man in the world. *L'Africain* was analysed as a text reflecting the existential crisis of the individual through an autobiographical narrative, while *Desert* was used to analyse the existential conflict between civilisation and nature. In *La quarantaine*, Le Clézio conceptualises isolation not merely as a physical condition but as an existential state characterised by temporal suspension and uncertainty.

The analysis of Modiano's works included a review of the novels *Un Pedigree* (2005), *La Place de l'étoile* (1996) and *Rue des Boutiques Obscures* (1978). These texts were selected because of their thematic focus on issues of memory, forgetting and the search for identity, which made them particularly significant in the context of existentialist literature. *Un Pedigree* was included in the study as an example of autobiographical prose in which existential motifs were expressed through the reconstruction of the past. In *La Place de l'étoile*, the narrative explores the problem of historical trauma and fractured identity, linking personal memory with collective experiences of the Second World War. *Rue des Boutiques Obscures* foregrounds oblivion and identity loss as structural principles of the narrative, where amnesia functions not only as a plot device but also as a metaphor for existential disorientation.

A comparative analysis of the works of Le Clézio and Modiano with existential concepts revealed how philosophical ideas were transformed in literary texts. The methodological approach provided a holistic consideration of the problem under study, combining philosophical, literary and comparative analysis. This not only identified the main existential motifs in the works of the two authors, but also determined their place in the development of French literature of the 20th and 21st centuries.

To ensure methodological consistency, the comparative analysis was conducted according to clearly defined criteria. These included:

- dominant existential themes (identity crisis, alienation, memory, and authenticity);
- narrative strategies (fragmentation, autobiographical elements, temporal discontinuity);
- spatial symbolism (nature versus urban space);
- the role of memory in the construction of identity;
- socio-historical context shaping existential experience.

This framework made it possible to identify both convergences and divergences in the artistic interpretation of existential motifs in the works of Le Clézio and Modiano.

3. Results

3.1. Existentialism as a philosophical and literary movement

Existentialism was formed as a philosophical trend focused on the study of the uniqueness of human existence, its freedom, responsibility and inevitable finitude. In contrast to rationalist concepts, it did not seek to create universal theories but focused on the subjective experience of the individual, ability to make meaningful choices and overcome the absurdity of existence. The central issues of existential thought were the problem of freedom, anxiety over endless possibilities of choice, the search for true existence and the awareness of the finitude of life.

The development of existentialist philosophy was influenced by the phenomenology proposed in the early 20th century. Reflections on being the space of human existence and its authenticity played a special role in its development (Efremov, 2026). Developing these ideas, Heidegger (2010) addressed the nature of human existence, denoting it using the term Dasein, emphasising its fundamental abandonment to the world. The study noted the awareness of personal mortality as the only way to transcend the everyday and exist authentically. If a person followed the predefined social norms, a state of inauthentic existence would be present, dissolving into an anonymous "They".

The relationship between crises and the awareness of individual existence has become a central theme of Jaspers (1971). The study argued that authentic existence is possible only in situations when a person is faced with insoluble problems of existence: fear, guilt, suffering and the inevitability of death. At such moments, the individual was freed from the usual patterns of perception, fulfilled the responsibility and moved on to true existence. These ideas were developed in philosophical works, emphasising that a personality is formed not through external circumstances, but through the way they are perceived.

Focusing on the analysis of human freedom, Sartre (2007) in *Existentialism is a Humanism* formulated the principle of "existence precedes essence", according to which a person does not have an initially given nature, but forms through personal choice. Such freedom, following the study, was absolute and inevitable, but at the same time carried the burden of full responsibility. The study examined the concept of "bad faith", describing it as a way of avoiding responsibility through self-deception, when a person shifts the choice to external circumstances or social norms. Freedom was perceived not as a privilege but as a burden that caused existential anxiety, since any decision formed a unique trajectory of existence without the possibility of relying on objective guidelines.

The question of the meaning of life and the human reaction to its absence became central to the reflections of Camus (2013), analysing the concept of absurdity. The essays argued that a person's desire to find the objective meaning of life is confronted with the silence of a world in which there are no predefined meanings. However, the awareness of this incommensurability should not lead to despair or flight into

illusion. Camus proposed an alternative way of existential rebellion, which consists of accepting life without false hopes and actively resisting the absurdity of existence.

The key categories of existentialist thought, such as absurdity, freedom, responsibility, loneliness, and the authenticity of existence, were reflected in philosophical works and influenced the literature of the 20th century (Figure 1).

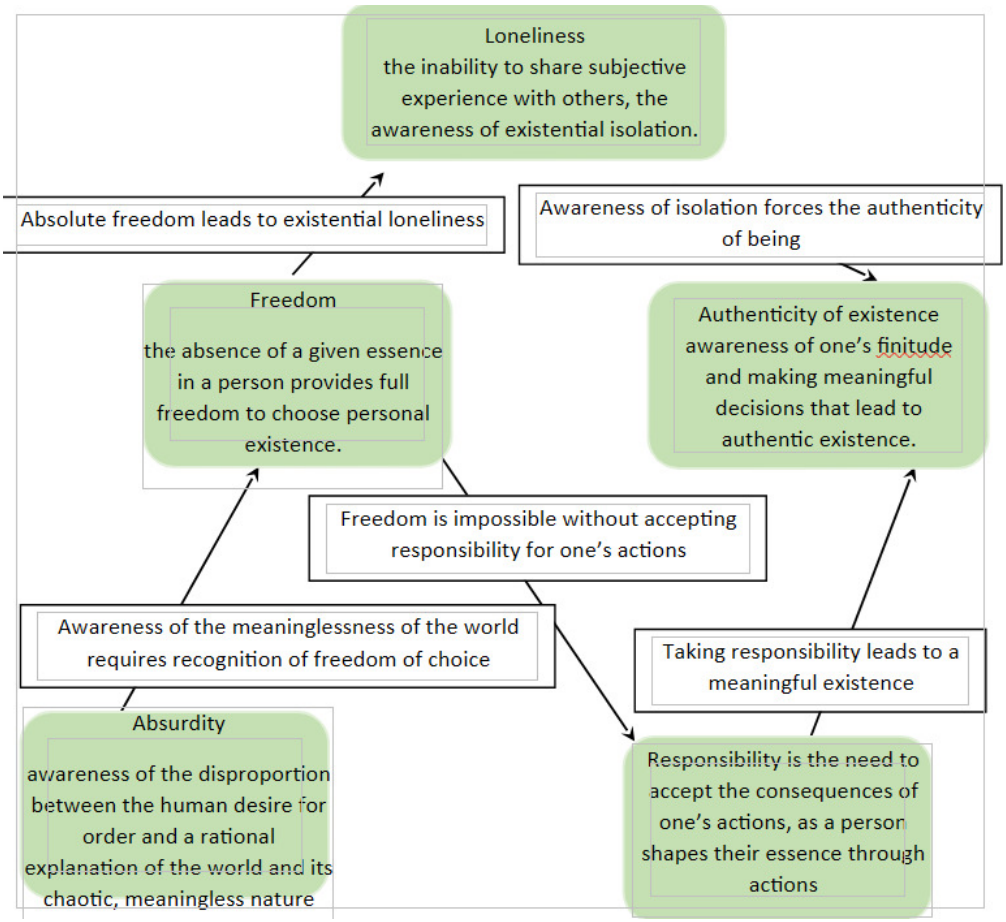


Figure 1. Interrelation of the categories of existentialism. Source: compiled by the author based on Haque et al. (2022), Davidov and Russo-Netzer (2022), Wharne (2023).

Figure 1 visualises the interdependence of existential categories, while their literary significance becomes evident in the ways narrative structures transform philosophical concepts into experiential and symbolic forms. In this sense, existential motifs emerge not as static thematic units but as relational processes embedded in narrative logic, character dynamics, and spatial configurations.

Thus, existentialism was formed as a movement that explores the inner world of a human, freedom and the ability to make meaningful choices. The philosophical ideas of Heidegger, Jaspers, Sartre and Camus laid the foundation for further artistic comprehension of these themes in literature, where they were reinterpreted through images, symbolism and narrative strategies.

3.2. Existentialism in literature

Existentialism significantly influenced the development of 20th-century literature, forming a unique genre, stylistic features and thematic trends. Literary works reflecting existential ideas sought not only to express philosophical concepts, but also to explore the inner world of the individual, their relationship to freedom, responsibility, absurdity, and loneliness. In contrast to traditional narrative forms, existential literature focused on the subjective perception of reality, the internal conflict of the individual and deep reflections on human existence (Chekal, 2020; Truba et al., 2024). The formation of existential literature in the 20th century was closely interconnected with the philosophical trends of the period. While in classical literature there were individual motifs of existential crisis, in modernist and postmodernist works, they became a central theme. The literary comprehension of existentialism contributed to the artistic interpretation of philosophical concepts such as the authenticity of existence, anxiety about freedom, and the search for the meaning of life (Shahab et al., 2020; Tukhtarova et al., 2021b).

Existential literature is characterised by the protagonist's inner conflict related to the awareness of freedom and the lack of objective guidelines, the use of stream of consciousness to convey chaotic thoughts, emotions and memories, and the theme of absurdity, reflecting the disproportion between the desire for order and the irrationality of the world, the motif of alienation, manifested in social isolation and awareness of fundamental loneliness, as well as the fragmentation of the narrative, expressed in the non-linear structure of the plot and the discontinuity of time layers (Table 1).

Existential literature formed special principles of narrative, distinguished by its focus on subjective perception, internal conflict, lack of a clear plot structure and emphasis on the concept of absurdity. It became not just an artistic reflection of philosophical ideas, but also an independent trend that rethinks the fundamental questions of human existence.

3.3. Existential motifs in the works of Le Clézio and Modiano

Existential motifs are central in the works of Le Clézio and Modiano, shaping the depth of their artistic comprehension of human existence. Their texts explore the themes of loneliness, alienation, identity crisis, memory and oblivion, presenting characters forced to face the uncertainty and absurdity of the surrounding world. The authors use autobiographical elements, metaphorical images and special narrative strategies, thereby conveying the subjective experience of the characters and their existential struggle.

Loneliness and alienation are key motifs in the works of Le Clézio, reflecting the existential search for the individual, the crisis of identity and attempts to comprehend one's place in the world. The characters find themselves in conditions of internal or external isolation, which prompts them to self-reflection and reassessment of their essence. These themes are expressed through autobiographical elements, symbolism of nature and the cultural contrast between traditional and modern societies.

Characteristic	Description	Impact on reader perception
The hero's internal conflict	The hero realises complete freedom and the absence of objective guidelines, which leads to existential uncertainty. Choices either lead to the search for authenticity or to a sense of meaninglessness of existence.	Creates tension, thoughts of the nature of choice and its consequences
Stream of consciousness	A way of conveying a character's inner world through chaotic thoughts, emotions and memories. Emphasises the uncertainty of thinking and the subjectivity of perceiving reality	Further demonstrates the mind of the character, creating a sense of involvement.
The theme of the absurd	The world is portrayed as irrational and unpredictable, with characters trying to find meaning in the face of uncertainty. There is often a gap between the usual perceptions of reality and its chaotic nature.	Creates a sense of anxiety, uncertainty and illogic in the surrounding world
Motive for alienation	The hero has difficulty establishing genuine connections with the world and the people around them. Alienation can be expressed in social isolation or awareness of the fundamental loneliness of existence.	Evokes a sense of isolation and emotional separation from the protagonist
Fragmented narrative	A rejection of the traditional linear plot, emphasising the fragmentation of perception. Past, present and future intersect to create a multi-layered narrative.	Creates the effect of a non-linear perception of time, and requires active participation in the reconstruction of the plot

Table 1. Characteristics of existential literature. Source: compiled by the author based on Shahab et al. (2020), Rumianowska (2022).

In *L'Africain* (Le Clézio, 2010), the protagonist's loneliness and alienation are shaped by a rupture with roots, fragmented identity, and the need to reconstruct the past as a key element of self-knowledge. The story is told in the first person, which enhances the effect of subjective perception, thereby deepening the portrayal of the protagonist's experience. Le Clézio reflects on childhood in Africa and the impact of this experience on personality. The protagonist is described as an outsider in both African and European culture, which creates an internal conflict. Existential motifs are manifested in reflections on the impossibility of belonging to a particular culture

or community. Space is substantial in this: African landscapes become a symbol of freedom, but at the same time emphasise the isolation of the protagonist, who cannot find himself in the existing coordinate system. The narrative structure is fragmented, reflecting the intermittent nature of memories and their impact on self-identity. Le Clézio uses descriptions of nature as a means of conveying existential loneliness: "Le silence de la savane me remplissait de crainte et d'admiration, comme si jétais dans un lieu n'appartenant à personne" ("The silence of the savannah filled me with fear and awe, as if I were in a place that belonged to no one"). Such a perception of the world demonstrates alienation and awareness of the impossibility of a full merger with this space.

A similar existential tension is articulated in *Désert* (Le Clézio, 2010), where loneliness emerges from the contrast between the natural and urbanised worlds and the erosion of traditional cultural values. The work focuses on two storylines: the story of a young woman, Lalla, who lives in harmony with nature, and the fate of ancestors, the Tuareg nomads, who were forced to leave their land under pressure from colonialists. Le Clézio conveys the existential conflict of the characters through the symbolism of the desert, which becomes a metaphor for loneliness and the endless space of freedom. Lalla perceives the desert as a home, but is forced to leave it and move to a modern city, encountering the impersonality of society. The alienation is emphasised through the language of the narrative: in the desert, life is described through sensual images, while in the city, the narrative becomes cold, laconic and mechanical. The existential choice in *Désert* is between an authentic but difficult life in the desert and the illusory stability of the civilised world. Lalla realises that existence in the city lacks authenticity, losing touch with nature and the roots. This conflict expresses the existential idea of the need to choose between authentic and imposed existence, which is a central theme of Sartre's existentialist framework.

In the novel *La Quarantaine* (Le Clézio, 1995), isolation is conceptualised as a form of spatial and existential suspension, where waiting and uncertainty function as conditions for self-reflection. Forced immobility intensifies internal conflict and prompts a reassessment of identity and memory, thereby transforming physical confinement into an existential experience. Literary techniques such as slow-motion narration and repeated images of the sea and fog create an atmosphere of suspension, in which the characters find themselves outside their usual temporal and social context. Leon, experiencing isolation, realises the relativity of all social attitudes. Within ordinary social conditions, Leon is faced with the need for reinternalisation. Waiting becomes not only physical, but also existential: the character loses previous landmarks, and existence becomes filled with uncertainty. The symbolism of the island reinforces the theme of alienation: it is a space where the boundaries between reality and illusion are blurred, which emphasises the existential crisis of the protagonist. The motif of quarantine in *La Quarantaine* correlates with the idea of existential anxiety described by Heidegger (2010), where the awareness of the finitude of existence makes the world more unpredictable and uncertain.

Loneliness and alienation in the works of Le Clézio are manifested through the themes of identity, cultural divide, the confrontation between nature and civilisation,

and forced isolation. Le Clézio uses autobiographical elements, the symbolism of natural spaces and slow motion to convey a state of existential uncertainty. The characters face a crisis of identity, the realisation that it is impossible to fully merge with the world and the need to choose between authentic and imposed existence. These motifs make Le Clézio's works a vivid example of artistic comprehension of existential ideas.

Loneliness and alienation are among the central motifs in Modiano's works, which are explored through the themes of memory, oblivion, identity and historical trauma. The characters of the novels are often in a state of existential crisis caused by the loss of the past, the impossibility of restoring biography and the realisation of alienation in the world. The narrative structures of texts reflect this uncertainty through the fragmentation of the plot, the disjointed timelines and the ambiguity of the characters.

In *Un Pedigree* (Modiano, 2005), personal history is reconstructed through the prism of memory, revealing the instability of identity and the fragmentation of autobiographical experience. In this text, the existential crisis is expressed in the protagonist's inability to fully reconstruct their biography: memories are fragmented, vague and filled with gaps, making identity unstable. The theme of loneliness is manifested in the protagonist's awareness of isolation from the past. The narrative constructs memory as an unstable epistemological framework, thereby problematising the possibility of coherent self-identification. The work is filled with descriptions of places that once mattered, but have now lost their connection to life: "Je n'ai jamais su si ces lieux avaient vraiment existé ou si je les avais seulement rêvés" ("I never knew whether these places really existed or if I had just made them up"). Modiano uses recurring motifs of forgotten streets, disappearing faces and indistinguishable shadows to convey a sense of loss and the impossibility of regaining what has been lost. The loneliness of the protagonist is reinforced by the narrative structure: the text is constructed as a series of fragments that do not add up to a coherent picture. Thus, the novel reflects the existential idea of the impossibility of full self-awareness through memory, emphasising the illusory nature of attempts to recreate one's past.

Modiano's *La Place de l'étoile* (1996) foregrounds loneliness as a consequence of historical trauma and fragmented identity within the context of post-war France. The novel deals with the problem of Jewish identity in France during and after the Second World War. The protagonist, Raphaël Schlömowitz, experiences an existential crisis caused by the realisation of alienation. The character feels torn between different identities, belonging to neither world. Existential alienation is manifested in the fact that the hero changes names and roles, trying to hide their identity, but this does not lead to integration into society. Raphaël is constantly searching for a place, but each choice only increases the sense of alienation. Modiano uses an ironic and fragmented narrative style to emphasise the absurdity of the situation in which Raphaël is present. Historical trauma becomes one of the mechanisms for the formation of existential loneliness. The past is imposed on the protagonist as an inevitable part of their identity, which cannot be removed, despite all attempts. Modiano shows that the alienation of the protagonists is caused not only by personal but also by collective historical circumstances that shape their existence in the space of traumatic memory.

In *Rue des Boutiques Obscures* (Modiano, 1978), identity is constructed as an open-ended process shaped by memory loss and narrative discontinuity. The protagonist, detective Guy Roland, suffers from amnesia and tries to recover identity, but as more becomes known, the uncertainty of existence is present. Memory loss functions as a structural metaphor, but also a way of artistic embodiment of existential loneliness. The protagonist is faced with the impossibility of objectively restoring the past and thus identity. The narrative is based on fragmentary memories, disparate details that do not add up to a single picture: "Chaque visage que je retrouvais dans mes souvenirs semblait se dissoudre dès que j'essayais de l'approcher" ("Every face I remembered dissolved as soon as I tried to approach it"). Existential motifs are manifested in the motif of a journey that does not lead to an ultimate goal. Guy Roland wanders endlessly around Paris, talking to people who only partially remember the past. This emphasises the impossibility of fully knowing oneself and gaining inner integrity. Modiano uses a monotonous, rhythmic narrative style that creates a sense of the protagonist's slow dissolution into the environment. The city in the novel becomes a labyrinth in which the individual loses boundaries, and the search for self is an endless process. In the end, the protagonist realises that existence has no clear basis, which resonates with the concept of the absurd as articulated in *The Myth of Sisyphus* by Camus (2013).

Loneliness and alienation in Modiano's works are revealed through the themes of memory, identity and historical trauma. The characters are in a situation of losing their past, being unable to integrate into society and facing the irrationality of personal existence. The author uses a fragmented structure, blurred boundaries of reality and recurring motifs of disappearance to convey an existential crisis. The texts show that loneliness is not only a social but also a philosophical condition associated with the inability to achieve a holistic self-perception.

Identity crisis, loneliness, and alienation are central themes in the works of Le Clézio and Modiano, but the authors interpret them in different ways, using different artistic techniques and focusing on different aspects of human existence. Le Clézio explores existential issues through human interaction with nature, the loss of cultural roots and the confrontation between civilisation and traditional ways of life. The characters experience a sense of alienation as a result of a break with historical and ethnic identity. In contrast, Modiano focuses on the loss of memory, the fragmentation of the past and the impossibility of restoring the integrity of the individual. The characters exist in a world of uncertainty, where the boundaries between real and fictional are blurred, and their biographies are subject to constant revision.

The differences in narrative strategies in the works of Le Clézio and Modiano reflect distinct modes of representing existential experience. In Le Clézio's prose, nature functions not merely as a recurring motif but as a structuring element of existential meaning. Natural space is associated with authenticity and the possibility of experiencing identity beyond the constraints of social and cultural norms. Through landscapes, journeys, and encounters with primordial environments, Le Clézio's characters articulate questions of belonging, freedom, and selfhood, which positions nature

as a medium for exploring existential authenticity rather than as a purely descriptive background.

Modiano's narrative strategy is grounded in a different spatial logic. In his works, urban space operates as a narrative framework through which existential instability is expressed. The city is represented as a fragmented and historically saturated environment, where anonymity, discontinuity, and unresolved memories shape the characters' experience of alienation. Instead of enabling self-discovery, the urban setting exposes the difficulty of reconstructing a coherent identity, as individuals are embedded in networks of social and historical traces that remain incomplete and opaque. In this sense, the city becomes a central means of articulating existential uncertainty and disorientation.

The opposition between natural and urban space thus reveals two contrasting models of existential representation. Le Clézio's texts foreground the search for authenticity through engagement with non-urban and pre-modern forms of experience, whereas Modiano's narratives emphasise the fragility of identity within modern historical and social conditions. This contrast demonstrates that existential motifs in contemporary literature are mediated by diverse spatial and narrative configurations rather than by a single aesthetic paradigm.

Table 2 presents the similarities and differences in the representation of existential ideas in the works of Le Cl'ezio and Modiano.

The analysis shows that both authors view loneliness and alienation as fundamental aspects of human existence, but with different emphases. Le Clézio uses the metaphor of travelling and interacting with nature as a way of finding true identity. The characters are forced to comprehend their existence through exile, loss of roots, and contact with civilisation. Modiano, in turn, shows the crisis of identity through the motif of disappearance, forgotten names, blurred memories and missing fragments of biography. The characters exist in the labyrinth of the past, where the search for lost history does not provide definitive answers. Thus, the similarities between the authors are manifested in a common existential problematic – the comprehension of loneliness and loss of identity, while the differences lie in the artistic means and interpretation of these themes. Le Clézio's work is built around a dialogue with nature and tradition, while Modiano conveys existential uncertainty through the themes of memory, oblivion and the urban environment.

Le Clézio and Modiano demonstrated the development of existential ideas in new conditions. While in classical existentialism, the crisis was viewed through inner freedom and choice, in these authors' works, it is associated with globalisation, historical memory and blurred identity. Their protagonists are not just restoring self-image, but are faced with the fragmentation of personal existence, which becomes dependent on external circumstances. This reflects a new phase of existential literature, in which loneliness and alienation are no longer limited to personal perception but are conditioned by the dynamics of the modern world.

Comparison criteria	Le Clézio	Modiano
The main theme of the existential crisis	Finding identity through connection to nature and cultural roots	An identity crisis caused by loss of memory and loss of the past
Narrative form	Lyrical narration, rich in descriptions, uses of nature symbols	Fragmented narrative, detective search, unreliable narrator
Reflection of loneliness and alienation	The characters experience alienation due to the loss of cultural roots and a break with tradition	The characters experience existential loneliness as their past is erased or blurred
The impact of memory on identity	The past is comprehended through landscapes, travel, and connection with the land	The past is reconstructed through scattered memories, archival documents, and stories of other people
Spatial organisation	The opposition of nature and civilisation, the importance of geographical space	Urban space, the labyrinth of streets, is a symbol of oblivion and uncertainty
Social and historical context	A cosmopolitan view, a study of the interaction of different cultures	Focus on historical memory, the impact of the Second World War and the traumas of the past

Table 2. Similarities and differences in the reflection of existential ideas by Le Clézio and Modiano. Source: compiled by the author based on Sterling (2022), Sopcak et al. (2022).

4. Discussion

The results of this study confirmed that the identity crisis is one of the key characteristics of existential discourse in contemporary French literature. The literary texts demonstrated that the sense of loss of personal integrity and the gap between the past and the present led the characters to search for meaning in the face of uncertainty. This conclusion correlated with the work of Löwith (2022), emphasising that Heidegger's existential philosophy defined identity through the awareness of the finitude of existence and the need to exist in conditions of abandonment in the world. In contrast to philosophical analysis, literary works conveyed the identity crisis through narrative strategies: fragmented narrative, blurred memory, and the inability to restore the continuity of biography.

This study also showed that existential anxiety arose in situations of crisis, forcing the characters to rethink their existence. This aspect was considered by Arafah and Nur (2024) in an analysis of the concept of liminal situations by Jaspers' concept of liminal situations, noting that the awareness of existence occurs in moments when a

person is faced with fundamental uncertainty. The similarity with the results of the present study was that the crisis state became a catalyst for change, but in the literary context, the existential choice often remained incomplete or ambiguous, while in the philosophical approach, it was viewed as a way to achieve authentic existence.

The study demonstrated that freedom was perceived as a fundamental but simultaneously disturbing aspect of existence. In literary texts, it was expressed in the absence of external reference points, which intensified feelings of loneliness and uncertainty. In the study by Van Dung (2020), existential freedom was considered in the context of responsibility, where choice inevitably led to the formation of personality. In contrast, in literary texts, freedom was not always perceived as a constructive force: characters often proved incapable of taking responsibility for personal existence, leading to a state of existential crisis.

A substantial role in the interpretation of existential motifs in the literature was played by memory, which became a tool of self-identification and at the same time a source of anxiety (Furman, 2023; Messina, 2025). The study by Wertz (2023) emphasised that existential psychology considers memory as the basis of personality formation, especially in the context of interaction with historical and social processes. This study confirmed the connection between memory and identity, but in literary texts, the emphasis was shifted to incomplete memories, their distortion and fragmentation, which led to the destabilisation of the individual, while in the psychological approach, memory was seen as a means of self-development.

The concept of absurdity has been one of the central aspects of the analysis of existential motifs in literature (Kychkyruk & Haluk, 2023). The present study demonstrated that the encounter with the absurdity of existence did not always lead to resistance or rebellion but was often expressed in the acceptance of uncertainty. This finding correlated with the study by Aboussaouira (2021), and Arinze and Onwuatuegwu (2020) in an analysis of the philosophy of Camus. Their research emphasised that absurdity is a fundamental property of human existence that requires awareness and acceptance. In fiction, in contrast to philosophical analysis, the reaction to absurdity varied: characters could either strive to overcome it or remain in a state of existential alienation.

Another substantial aspect of this research was the study of the social nature of existential problems. In the literature, loneliness and alienation have often been presented as personal experiences associated with the loss of identity and a break with the past (Lytvyn et al., 2018). Baert et al. (2022) analysed existentialism in the context of social theory, and the authors concluded that modern conceptions of freedom, identity, and responsibility have changed under the influence of social structures. In contrast, the literary texts showed that the loneliness of the characters was not so much a social phenomenon as the result of their internal contradictions and inability to integrate into the surrounding world.

The study highlighted the symbolic interpretation of existential motifs, which correlated with the analysis of mythological elements in contemporary literature presented in O'Brien (2024). In the philosophical context, myth was seen as a way to

comprehend existential problems, while in fiction it acquired a different function, becoming a tool for constructing a disjointed narrative and a way to convey the fragmentation of human experience. While the researcher's mythological analysis focused on universal structures, this study emphasises the individual interpretation of mythological consciousness in literature.

Phillips (2020), in a study of contemporary trends in French literature, emphasised that since the 2000s, the revision of historical memory, cultural fragmentation and the impact of globalisation on personal identity have become key themes. This study determined that the works of Le Clézio and Modiano do reflect these processes, demonstrating the instability of identity in a rapidly changing world. However, in contrast to the researcher of literary criticism in a broad context, this study focuses on existential uncertainty and its artistic manifestations.

The analysis of Le Clézio's works determined that the texts often link existential issues to travel and the attempt to find roots through interaction with the environment. Kpoumié (2020) also examined the theme of travel and nostalgia in the novels, emphasising that in Le Clézio, the characters' journeys serve not only as physical movement but also as a way of making sense of their existence. In this aspect, the results are similar, but the present study additionally notes that the process of travelling does not lead to final self-determination: Le Clézio's characters remain in a state of existential uncertainty, while the researcher focuses on their interaction with cultural contexts.

The problem of identity raised in this study is reflected by Özgüler (2022) in an analysis of how Le Clézio's perception of personality is formed through the confrontation between the past and the present. The study concluded that characters experience a gap between personal memories, family history and the surrounding world, which leads to an existential crisis. The present study confirms this conclusion but also establishes that the identity of Le Clézio's characters is unstable not only because of the influence of the past, but also because of the awareness of the inevitable loss of authenticity in the modern world. Thus, the difference between the approaches is that the researcher considers the problem of identity mainly in the socio-cultural context, while this study emphasises the philosophical aspect of existential uncertainty.

Existential loneliness and its connection with the environment were also analysed by Robertson (2024) in a literary review of *Désert*, noting that the desert in the work is not just a landscape, but a symbol of inner emptiness and reflections on the meaning of life. The present study confirmed this conclusion, showing that the space of nature in Le Clézio is an existential contrast between chaotic civilisation and secluded existence. However, in contrast to the researcher, who considered the symbolic aspect of the landscape primarily in terms of its aesthetic function, this study focuses on the subjective perception of space by the characters, which intensifies their existential conflict.

Historical memory and its influence on the formation of existential motifs were identified as a key element of Modiano's works. Gunderson (2022) also explored this topic, considering how the structure of memories in texts shapes narrative strategies.

The study argues that the detective structure of Modiano's novels reflects the process of searching for identity through fragmentary memories, which coincides with the results of the present study. However, the difference between the approaches is that the researcher's work focuses on the cognitive aspects of memory, while this study analyses how oblivion and instability of the past affect the existential state of the characters.

The study demonstrated that disappearance and oblivion are central in Modiano's works, enhancing the uncertainty of characters. Goloborodko and Lepetiukha (2020) noted that in later novels, the theme of loss of reality is increasingly emphasised, and the narrative becomes even more fragmented. The present study confirmed this conclusion but also found that the uncertainty in Modiano's texts is not only related to the loss of memory, but also to the awareness of the absurdity of human existence. While the researchers analysed the evolution of Modiano's style, this study focuses on the philosophical significance of these changes.

Existentialism in contemporary literature is often reinterpreted in the light of new social and cultural challenges (Tukhtarova et al., 2021a). Bondar (2022) examined this process by comparing classical existentialism with its transformation in the 21st century. The study noted that contemporary writers are rethinking traditional existential ideas, combining them with the themes of globalisation and historical memory. This study confirmed this hypothesis, finding that in the works of Le Clézio and Modiano, the classic motifs of loneliness, absurdity and identity are combined with more complex narrative structures that reflect a changing perception of existential problems. The difference, however, is that the study addressed the sociological aspect of change, while this study analyses the artistic interpretation of existential states.

Comparison of the research results with existing works confirmed the key provisions of existential philosophy but showed that these concepts received a specific artistic embodiment in literature. There were commonalities in the issues of identity, crisis of existence, memory and absurdity, but the differences concerned the manner of their expression: philosophical studies offered a systematic analysis of existential problems, while literary texts conveyed subjective experience, reflecting the personal feelings of the characters in conditions of uncertainty.

5. Conclusions

The study of existentialism in philosophy and literature identified the key concepts of this trend and analysed their artistic embodiment in the works of Le Clézio and Modiano. Existentialism developed as a philosophical response to the crisis of identity, freedom and the meaning of human existence, and its main categories of absurdity, loneliness, responsibility and authenticity are reflected in literature. The analysis has shown that Le Clézio and Modiano, relying on existential ideas, adapted them to modern social and historical realities, changing approaches to the interpretation of loneliness and alienation.

Le Clézio's work demonstrated the connection between existential issues and questions of cultural identity, loss of roots and human interaction with the surrounding world. The characters faced alienation caused by the gap between tradition and modernity, the inability to fully integrate into the civilised world and the loss of connection with the natural environment. Existential motifs were manifested through the awareness of loneliness not only on a personal level, but also in a broader cultural and historical context. Le Clézio's narrative strategies included detailed descriptions of nature, symbolism of the path and the confrontation between urban and natural space, which emphasised the existential conflict of characters.

In Modiano's works, existential ideas were presented through the motifs of memory, oblivion and identity instability. The characters existed in a state of uncertainty caused by the blurring of their past, the impossibility of recovering their personal history and the feeling of alienation in the world. Alienation was manifested not through confrontation with external conditions, but through an internal crisis associated with the awareness of personal instability. The narrative was fragmented, and the main characters were in constant search of the missing elements of their biography, which made their existence dependent on memories and interpretations of the past.

A comparative analysis demonstrated that both authors addressed the theme of existential loneliness, but with different emphases. Le Clézio examined the identity crisis in the context of globalisation, the post-colonial world and the loss of traditional cultural values, while Modiano focused on the impact of historical memory and collective trauma on personal existence. Whereas Le Clézio's search for identity was through interaction with the environment and perception of personal roots, Modiano's was through attempts to recover the lost past, but this process was fragmented and endless.

The study has confirmed that existentialism remains a substantial trend in contemporary literature, transforming depending on historical and cultural changes. The works of Le Clézio and Modiano demonstrated the evolution of existential motifs, adapting them to the realities of the modern world. Their works have shown that loneliness and alienation can be caused not only by personal choice, but also by the influence of historical memory, global processes and identity instability.

References

Aboussaouira, R. (2021). *Albert Camus' "The myth of Sisyphus": Absurdism and the question of happiness*. Hassan II University of Casablanca.

Arafah, Y., & Nur, M. Y. C. (2024). Modern human concept philosophical perspective Karl Jaspers. *Humanist: As'adiyah International Journal of Humanities and Education, 1*(2), 85–99.

Arinze, A. T., & Onwuatuegwu, I. N. (2020). The notion of absurdity and meaning of life in Albert Camus existentialism. *Open Journal of Philosophy, 10*(4), 528–538. https://doi.org/10.4236/ojpp.2020.104037

Baert, P., Morgan, M., & Ushiyama, R. (2022). Existence theory: Outline for a theory of social behaviour. *Journal of Classical Sociology, 22*(1), 7–29. https://doi.org/10.1177/1468795X21998247

Bondar, O. (2022). Between existentialism and anti-existentialism. *Wisdom, 21*(1), 204–216. https://doi.org/10.24234/wisdom.v21i1.605

Camus, A. (2013). *The myth of Sisyphus*. Penguin Books.

Chekal, L. (2020). Consciousness and self-consciousness in psychoanalytical, phenomenological and existential discourse. *Humanities Studios: Pedagogy, Psychology, Philosophy, 8*(4), 130–137. https://doi.org/10.31548/hspedagog2020.04.130

Cropper, T., & Browne, J. (2024). The existentialist dilemma: Analyzing nihilism, anxiety, and self-identity in French philosophy at the end of the 20th century. *Studies in Social Science & Humanities, 3*(1), 1–9. http://doi.org/10.56397/SSSH.2024.01.01

Davidov, J., & Russo-Netzer, P. (2022). *Existential authenticity: Theory and practice*. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-07842-2

Efremov, A. (2026). Age-Specific Mental Health Profiles of Combat Veterans: Post-Traumatic Stress Disorder and Related Disorders. *Journal of Rational-Emotive and Cognitive-Behavior Therapy, 44*(1), 4. https://doi.org/10.1007/s10942-025-00637-7

Furman, A. (2023). Methodology and psychology in reflective architecture of methodology concenters. *Psychology and Personality, 13*(2), 9–83. https://doi.org/10.33989/2226-4078.2023.2.288169

Girard, E. (2024). Existentialism and individualism: A study of its influence on French society. *Journal of Philosophy, Culture and Religion, 7*(2), 46–56. https://doi.org/10.47604/jpcr.2607

Goloborodko, I., & Lepetiukha, A. (2020). Virages des visions de Patrick Modiano. *Neophilologus, 104*(4), 165–176. https://doi.org/10.1007/s11061-019-09630-9

Gosetti-Ferencei, J. (2020). Existentialism as literature: The twentieth century. In J. Gosetti-Ferencei, *On Being and Becoming: An Existentialist Approach to Life* (pp. 88–106). Oxford Academic. https://doi.org/10.1093/oso/9780190913656.003.0006

Guignon, C. (2020). Existentialist ethics. In J. P. DeMarco & R. M. Fox (Eds.), *New Directions in Ethics: The Challenges in Applied Ethics* (pp. 73–91). Routledge. https://doi.org/10.4324/9781003031475

Gunderson, K. (2022). Memory, history, cognition: Using systems-theoretical social theory to illuminate Patrick Modiano's *enquêtes policiéres*. *History Notebooks, 38*(2), 46–72. https://doi.org/10.7202/1099404ar

Haque, M. A., & Reyaz, D. A. (2022). The existential predicament: Finding meaning in a meaningless world. *Creative Launcher, 7*(3), 101–108. https://doi.org/10.53032/tcl.2022.7.3.12

Heidegger, M. (2010). *Being and time*. State University of New York Press.

Jaspers, K. (1971). *Philosophy of existence*. University of Pennsylvania Press.

Kpoumié, M. N. (2020). Travel and nostalgia in *Onitsha and The African* by JMG Le Clézio. *Carnets, 20*, 1–13. https://doi.org/10.4000/carnets.12237

Kychkyruk, T., & Haluk, M.V. (2023). Peace through the prism of civilizations: Approaches, interactions and crises. *Humanities Studios: Pedagogy, Psychology, Philosophy, 11*(4), 196–204. https://doi.org/10.31548/hspedagog14(4).2023.196-204

Le Clézio, J. M. G. (1995). *La quarantaine*. Gallimard.

Le Clézio, J. M. G. (2010). *Desert*. Atlantic Books.

Le Clézio, J. M. G. (2010). *L'Africain*. Mercure De France.

Löwith, K. (2022). Heidegger: Problem and background of existentialism. In K. Löwith, *Karl Löwith: Heidegger – A Thinker in a Time of Crisis: Complete Writings* (pp. 102–123). J. B. Metzler Berlin. https://doi.org/10.1007/978-3-662-65943-4_5

Lytvyn, V., Vysotska, V., Burov, Y., Bobyk, I., & Ohirko, O. (2018). The linguometric approach for co-authoring author's style definition. In: *Proceedings of the 2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems, IDAACS-SWS 2018* (pp. 29–34). Lviv, Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/IDAACS-SWS.2018.8525741

Mahajan, A. (2021). Existentialism: A philosophical approach in literature. *Turkish Online Journal of Qualitative Inquiry, 12*(5), 5124–5133.

McElroy, D. D. (2022). *Existentialism and modern literature: An essay in existential criticism*. Open Road Media.

Messina, A. (2025). Dissociative symptoms in schizophrenia: a connection to inflammatory processes. *Middle East Current Psychiatry, 32*(1), 96. https://doi.org/10.1186/s43045-025-00593-z

Modiano, P. (1978). *Rue des boutiques obscures*. Gallimard.

Modiano, P. (1996). *La place de l'étoile*. Gallimard.

Modiano, P. (2005). *Un pedigree*. Gallimard.

O'Brien, M. D. (2024). *Mythological cognition in contemporary science fiction and fantasy*. University of Glasgow.

Özgüler, C. (2022). Identity problem in Le Clézio's novels. *RumeliDE Journal of Language and Literature Studies*, Ö11, 444–450. https://doi.org/10.29000/rumelide.1146668

Phillips, M. (2020). French studies: Literature, 2000 to the present day. *Year's Work in Modern Language Studies*, *80*(1), 209–260. https://doi.org/10.1163/22224297-08001010

Quindadiarto, F. (2023). Identity crisis and alienation in The Metamorphosis: Existential approach. *LITERA KULTURA: Journal of Literary and Cultural Studies*, *10*(2), 28–34. https://doi.org/10.26740/lk.v10i2.49425

Rahimipour, S. (2021). Comparison of existence concept in the works of Jean-Paul Sartre and Hemingway with emphasis on the educational views of existentialism. *Iranian Journal of Comparative Education*, *4*(3), 1317–1328. https://doi.org/10.22034/IJCE.2020.232864.1160

Robertson, H. L. (2024). Literary review of Into the Desert. *EON*, *5*(1), 18–23. https://doi.org/10.56177/eon.5.1.2024.art.3

Rumianowska, A. (2022). The experience of conflicts (internal and external) and the sense of meaning in life. An existential approach. *Pedagogical Forum*, *12*(2), 321–334. https://doi.org/10.21697/fp.2022.2.24

Sartre, J. P. (2007). *Existentialism is a humanism*. Yale University Press. https://doi.org/10.2307/j.ctv15vwkgx

Sasan, J. M. (2023). Existentialism and its influence on our understanding of knowledge, truth, morality, values, and religion. *European Journal of Learning on History and Social Sciences*, *1*(1), 40–48. http://dx.doi.org/10.61796/ejlhss.v1i1.11

Shahab, A., Faruk, F., & Rokhman, A. (2020). French literature: From realism to magical realism. *Poetics Journal*, *8*(2), 170–181. https://doi.org/10.22146/poetika.v8i2.58651

Sopcak, P., Kuiken, D., & Douglas, S. (2022). Existential reflection and morality. *Frontiers in Communication*, *7*, 991774. https://doi.org/10.3389/fcomm.2022.991774

Sterling, J. (2022). Reflections on existential types. *arXiv preprint arXiv:2210.00758*. https://doi.org/10.48550/arXiv.2210.00758

Truba, H., Khrapatyi, S., Harashchuk, K., Shvets, D., & Proskurnia, A. (2024). Psycholinguistic underpinnings of image formation: Suggestion and manipulation in the educational network discourse. *Thinking Skills and Creativity*, *52*, 101496. https://doi.org/10.1016/j.tsc.2024.101496

Tukhtarova, A. S., Issakova, S. S., Kussaiynova, Z. A., Kenzhemuratova, S. K., & Nassyritdinova, A. N. (2021a). Linguocultural approach to audiovisual translation on the example of "game of thrones". *International Journal of Society Culture and Language*, *9*(3), 49–63.

Tukhtarova, A. S., Issakova, S. S., Toxanbayeva, T. Zh., Aitbayeva, N. K., & Albekova, A. Sh. (2021b). Linguocultural deviations in russian translation of the novel "twilight" by S. Meyer. *International Journal of Society, Culture and Language*, *9*(2), 189–202.

Van Dung, V. (2020). The thought of human in existentialism. *European Journal of Human Resource Management Studies*, *4*(1), 72–84. http://dx.doi.org/10.5281/zenodo.3732228

Wertz, F. J. (2023). The history of humanistic and existential psychology: The possibility and cultural contexts of renewal in science. In L. Hoffman, L. X. Vallejos, D. Hocoy, P. Tummala-Narra, & E. DeRobertis (Eds.), *APA Handbook of Humanistic and Existential Psychology*. American Psychological Association.

Wharne, S. (2023). Existential relevance to everyday life. In C. Arnold-Baker, S. Wharne, N. H. Dowek, N. Gibson, J. Molle, & E. van Deurzen (Eds.), *Existential Therapy: Responses to Frequently Asked Questions* (pp. 130–160). Routledge. https://doi.org/10.4324/9781003355700

Zhou, R. (2024). Existentialism, authenticity and modern debates over incapable individuals. *Transactions on Social Science, Education and Humanities Research*, *11*, 469473. https://doi.org/10.62051/s63mtd56

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 312 – 327

THE FUNDAMENTAL PRINCIPLES, ARGUMENTS, AND IMPLICATIONS OF EXISTENTIALISM IN THE LIGHT OF SARTRE)

Gizem Şerifoğulları*

Abdülkadir Çüçen**

Abstract: This article first examines the historical and philosophical elements that prepared the ground for the emergence of existentialism. The individualization and search for freedom brought about by the modern era, along with the resulting social disappointments, confronted human beings with experiences of alienation and meaninglessness. In this context, the contributions of Kierkegaard, Nietzsche, Jaspers, Heidegger, and other existentialist thinkers are presented as a historical background, thereby clarifying Sartre's distinctive position within this tradition. The second part discusses the common assumptions of existentialism, focusing particularly on the critiques directed against metaphysical ontology, determinism, and universal moral laws. The third part analyzes Sartre's unique approach to existentialism in light of the criticisms directed at him, and through the concepts of freedom, nothingness, consciousness, and project. At this point, Sartre's principle that "existence precedes essence" is placed at the center, and the aspects of his conception of freedom that necessarily lead the human being toward responsibility and action are examined. The aim of the article is to challenge the frequently asserted interpretation of Sartrean existentialism as a doctrine of despair and passivity, and instead to reinterpret his philosophy on the basis of freedom, responsibility, and action, presenting it as a philosophy of humanism and hope.

Keywords: existentialism, nothingness, freedom, human, consciousness, project.

1. Introduction

With the advent of the modern era, the dominance of instrumental reason brought about a series of profound social transformations. In particular, liberalism –whose theoretical foundations in the seventeenth century emphasized individual rights and freedoms – initiated struggles for individuality and liberty. Yet these struggles did not yield the expected results. At the height of liberal thought in the eighteenth, nineteenth, and twentieth centuries, the promises made to the individual largely failed to materialize (Taylor, 1989). As Sartre himself observed, the human being, under the hegemony of instrumental reason and in pursuit of liberation, finds that rather than attaining freedom, one's choices are increasingly restricted (Şerifoğulları, 2023). The gains of the Industrial Revolution, rather than securing individual freedom, ultimately served the interests of the capitalist system. Through class divisions and the dominance of rational calculation and self-interest, the human being was reduced to the

*Department of Philosophy, Bursa Uludag University, Bursa, Turkey; ORCID: 0000-0002-6613-5333; gizemserifogullari@gmail.com ** Department of Philosophy, Bursa Uludag University, Bursa, Turkey; ORCID: 0000-0001-8092-1472; kadir@uludag.edu.tr

©2026 Şerifoğulları and Çüçen. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

status of an object, transformed into a mechanical entity alienated from its own existence. In this process, the pursuit of freedom was overshadowed by instrumentality grounded in self-interest, while human purposiveness was forgotten. By the twentieth century, the rise of technology and its use in military armament during the First and Second World Wars led to inevitable social and economic collapse (Flynn, 2006). Thus, the practical consequences of modernity's project, which had been centuries in the making, culminated in widespread human disappointment.

The problems generated by seventeenth-century rationalism, eighteenth-century Enlightenment, and nineteenth-century positivist and socialist-democratic thought confronted twentieth-century humanity with military, political, economic, and social crises on an unprecedented scale, producing what may be called an existentialist crisis. Instead of delivering on the ideal of progress and the promise of prosperity, this trajectory resulted in exploitation and the danger of domination. The restriction of human freedom and the experience of alienation increasingly provoked a questioning of existence that formed the basis for the emergence of existential philosophy. In this sense, the difficulties and estrangement experienced by modern humanity in a secular world, together with its attempts to overcome them, were articulated within the existentialist movement that came to the fore in twentieth-century philosophy. Existentialism, by taking seriously the individual's effort to secure a place in the world, thus presented itself as a philosophy of revolt against the reduction of the human being to a mere mechanical entity. From Kierkegaard's leap of faith to Nietzsche's revaluation of values, from Jaspers' search for transcendence to Heidegger's analysis of *Dasein*, many existentialist philosophers have addressed the finitude of human existence from different perspectives. Rather than sharing a single doctrinal principle, these approaches display a family resemblance in their insistence on returning philosophy to concrete existence, freedom, and the individual's confrontation with meaning, anxiety, and responsibility. Within this broad field, the formula "existence precedes essence" is most explicitly articulated by Sartre and becomes a guiding thesis for his distinct existentialist position. Sartre's originality, as developed systematically in *Being and Nothingness*, lies in grounding this thesis in a phenomenological ontology of consciousness, nothingness, and projects, thereby linking freedom to responsibility and action rather than to quietism or despair (Sartre, 1992; Sartre, 2007). What, then, is Sartre's unique contribution with regard to the fundamental principles, Arguments, and implications of existentialism? This article aims to demonstrate how Sartre reshapes these principles. Contrary to interpretations frequently encountered in the literature that portray existentialism as a philosophy of despair and passivity, the present study highlights how Sartre's existentialism can be understood as a philosophy of freedom, responsibility, and revolt – ultimately a philosophy of action and humanism. In this way, the study seeks to make a contribution to the existentialist literature by emphasizing Sartre's originality in contrast to his contemporaries. To this end, the article presents the philosophies of existentialist thinkers within a general framework and, in light of this, traces Sartre's analyses in terms of the principles and arguments of existentialism. In order to discern Sartre's distinctiveness, it is first necessary to examine existentialism in its historical development.

2. A historical overview of existentialism

Existentialism appears as a movement that invites human beings to recognize their own existence. In this respect, rather than attempting to understand the human being on a purely rational or absolute level, existentialism approaches the individual through lived experience and, in doing so, offers different perspectives on the question, "How should we live?" Although existential philosophy is often associated with nineteenth- and twentieth-century thinkers, certain philosophical tendencies anticipating existential concerns can be traced back to earlier reflections on human existence.

The natural philosophers, in their attachment to the physical world, neglected the inquiry into the meaning of life and thus left the human being in the background. For Socrates, the danger was that human beings, absorbed in their concern with objects, forgot to reflect on themselves. In this sense, Socrates stands out for directing philosophical inquiry toward the human being and human action. While it would not be historically accurate to describe Socrates as an existentialist, his emphasis on self-examination and the examined life prepared an important intellectual background for later existential thought.

In the Middle Ages, Augustine occupies an important place in the genealogy of existentialist thought with his reflections on free will and the problem of evil. Like the existentialists, Augustine resisted abstract and speculative doctrines, instead making his own life the object of reflection and thereby showing that philosophy cannot be separated from the practical life of the individual (Çüçen, 2018, pp. 60–61). In the seventeenth century, Pascal emphasized the limits of reason, arguing that the human being cannot be understood through the rational faculty alone and that intuition must also be engaged. In this respect, Pascal prepared the ground for the emergence of existentialist thought. Nevertheless, existentialism in its more explicit philosophical form emerged primarily in the nineteenth century, within the broader intellectual crisis that followed the decline of systematic idealist philosophies such as Hegel's (Löwith, 1964). As Löwith argues, the dissolution of the Hegelian system did not merely mark the end of a comprehensive metaphysical synthesis; it opened a problem-space in which the individual's concrete existence, finitude, and historical situation could no longer be subsumed under a reconciliatory totality (Löwith, 1964). In this context, existential thought emerges as a response to the experience that meaning, value, and freedom must be confronted at the level of lived life rather than secured by an overarching philosophical system.

Just as Socrates opposed the natural philosophers for neglecting the human being in favor of the physical world, Kierkegaard also rejected the Enlightenment view of the human as a mechanical entity, and instead laid the foundations for recognizing human existence. For him, the acknowledgment of human existence is possible only independently of science. Accordingly, Kierkegaard found the solution to this problem in turning toward God. Moreover, Kierkegaard developed a distinctive philosophical method known as "indirect communication," through which philosophical insights are expressed not only through systematic argument but also through literary forms such as pseudonymous writings and existential narratives. Nietzsche, by contrast, rejected metaphysics and developed an opposing discourse to Kierkegaard. Since he believed

that the human being should not be understood through metaphysical abstractions but within the framework of this world, "Nietzsche made his great choice in favor of the mortal world that Kierkegaard had rejected" (Blacham, 1967, p. 45). This is also the reason for his appeal through the character of Zarathustra: "I beseech you, my brothers, remain faithful to the earth and do not believe those who speak to you of otherworldly hopes!" (Nietzsche, 1969, p. 12). Nietzsche's exhortation constitutes a critique of the negative effects of religious dogmas on human life in his time. Indeed, adopting an atheistic perspective, Nietzsche sought to replace hopes in otherworldly realms with the possibility of becoming the Übermensch, grounded in the will to power within this world (Crowell, 2012).

The broader intellectual context of the nineteenth century also included reflections on alienation and the transformation of human life under modern social conditions. In particular, Marx's analysis of alienation drew attention to the ways in which modern economic and social structures could estrange individuals from their own activity and from their human essence (Marx, 2000; Marx, 1976). Although Marx is not generally classified as an existentialist philosopher, his analysis of alienation contributed to the intellectual environment in which questions of human freedom, responsibility, and authenticity became increasingly central.

Following Kierkegaard and Nietzsche, Jaspers claimed that philosophy could no longer remain on the same plane. Against the dominance of rationalism and the positivist tendency to transform philosophy into either science or metaphysics, Jaspers argued that with these two thinkers, philosophy had turned decisively toward the human being. As he put it, "Philosophy intensifies the human being, making him a participant in reality, and thus the human becomes himself" (Jaspers, 1957, p. 28). Whereas Nietzsche's philosophy was grounded in atheism and Kierkegaard's in Christianity, Jaspers advanced a philosophy of existence centered on the concept of Transcendence, which he identified with God. Influenced by Kierkegaard and Jaspers, Marcel argued that modern man, by reducing himself to the status of an object and thereby becoming alienated, had lost the possibility of authentic existence. For Marcel, to exist as a human being means to be in relation with other beings in the world and, at the very least, to live together with other human beings. In his view, alienation stems from the obsession with science and technology and from a failure to orient oneself toward God (West, 2010). Marcel develops these themes most explicitly in *The Metaphysical Journal*, where he places existence and lived experience at the center of philosophical reflection and thereby reinforces the existential tradition's emphasis on concrete, situated human life (Marcel, 1952). Marcel's reflections on lived experience are also significant because they illustrate the close relationship between philosophical reflection and literary expression within the existential tradition, a feature that later becomes particularly visible in Sartre's philosophical and literary works. Marcel also wrote plays, which further underscores the literary–philosophical nexus within the existential tradition. However, this literary–philosophical proximity becomes especially salient in Sartre: several central existentialist motifs are first articulated in his novels and plays, long before receiving their most systematic philosophical formulation in *Being and Nothingness* (Sartre, 1992). While Marcel articulates the existential

crisis through transcendence and intersubjective existence, other twentieth-century thinkers reframe the question of existence through phenomenological analyses of human being-in-the-world.

For this reason, the human being must be bound to God in order to realize itself. In Heidegger, we find the claim that human existence cannot be grasped through logical principles, articulated by the concept of *Dasein*, which abstracts from all previous determinations of the human being (Heidegger, 1962). Merleau-Ponty, on the other hand, directs his radical critique against Descartes' dualism, emphasizing that the division of the whole human into body and mind, and its reduction to a machine-like object, results in alienation. As he writes, the body is not an object for the mind; it is our means of communication with the World (Merleau Ponty, 1964a, p. 16). With this statement from *Eye and Mind*, he underlines the futility of mechanizing distinctions and highlights the body as the primary locus of human existence (Merleau Ponty, 1964b).

Simone de Beauvoir emphasizes that the human being is not a predetermined entity and thus, in objectifying itself, exercises freedom of choice. In this respect, she highlights the awareness of responsibility and thereby extends existentialism into the ethical domain (de Beauvoir, 1949). Another existential philosopher who did not remain indifferent to the effort of understanding human beings through reason is Camus. By emphasizing the irrationality between the human and the world, he argued that neither the human nor life itself can be rendered meaningful by reason (Camus, 1991). Yet, unlike his contemporaries, Sartre offers a distinctive grounding of what existentialism is. For Sartre, the conviction that the human being is necessarily free determines his understanding of existentialism and, in turn, directly shapes the fundamental principles, arguments, and orientations of existential philosophy (Sartre, 2007).

3. The common assumptions of existentialism

Existentialism is often divided into two broad categories: theistic and atheistic existentialism. Yet this division should not be understood as implying internal uniformity. Theistic existentialists such as Kierkegaard, Jaspers, and Marcel emphasize transcendence and the individual's relation to God, whereas atheistic existentialists such as Nietzsche, Heidegger, and Sartre interpret existence in terms of immanence and human freedom without appeal to the divine (Flynn, 2006). At the same time, this classification should be treated as a heuristic rather than a settled historiographical framework, since the self-understandings of major figures and the boundaries of the "existentialist" label remain contested in the literature (Macquarrie, 1972; Crowell, 2012). Still, neither group presents a completely unified outlook; significant differences remain within each tradition. Despite these divergences, however, it is possible to identify several common assumptions that constitute the foundation of existentialist thought. These "common assumptions" are best understood not as a single shared doctrine but as a family resemblance: recurring problem-fields – such as finitude, freedom, anxiety, and the critique of abstraction – that are developed in divergent and

sometimes incompatible ways (Crowell, 2012). Before highlighting Sartre's distinctive contribution, these shared features must be considered in detail.

At its core, existentialist philosophy rejects traditional metaphysical ontology. From Plato's realm of the Forms to Hegel's Absolute Spirit, the history of Western philosophy has often displaced the individual into abstract, universal categories, neglecting the immediacy of lived human experience. Existentialism insists that human existence cannot be adequately understood through logical deduction or speculative abstraction; instead, it must be approached in its concreteness, as it is lived (Crowell, 2012; Moran, 2000). In this respect, existentialism marks a radical break with modern philosophy's reliance on epistemology and science as the foundations of ontology, arguing instead for a philosophy grounded in human finitude and subjectivity. Nevertheless, this critique does not entail a uniform metaphysical position. For some thinkers, the point is to reorient philosophy toward existence through faith or transcendence (e.g., Kierkegaard, Marcel), whereas for others it is to articulate an immanent analysis of being-in-the-world or human reality without recourse to metaphysical guarantees (e.g., Heidegger, Sartre) (Flynn, 2006; Crowell, 2012). What unifies these approaches is the insistence that the meaning of "being human" must be approached from within existence itself, not imposed upon it by an external essence or a purely abstract system. This emphasis prepares the ground for Sartre's ontological claim that human reality lacks a pre-given essence and must be understood through the structures of consciousness and existence (Sartre, 1992). Another shared feature is the critique of universalist moral systems. Traditional ethics, based on transcendent or metaphysical principles, tends to impose uniform laws upon all individuals, disregarding human particularity. Existentialist thinkers consistently resist such moral abstraction, claiming that it obscures the lived reality of choice and responsibility. For Kierkegaard, authenticity arises in the individual's "leap of faith," while for Nietzsche, values must be created anew in response to the death of God. Sartre and de Beauvoir, in turn, advance ethical perspectives in which freedom entails an unavoidable responsibility toward oneself and others (Macquarrie, 1972; de Beauvoir, 1948). In each case, existentialists affirm that moral meaning emerges not from universal dictates but from the situated decisions of individuals. Yet it is important to note that "situated decision" does not imply ethical arbitrariness: existentialist critiques of universalism typically aim to recover the concreteness of moral life – its ambiguity, conflict, and responsibility – rather than to deny the possibility of ethical seriousness (West, 2010; Macquarrie, 1972). This point is central for Sartre, since many criticisms directed at his existentialism – quietism, relativism, and subjectivism – presuppose that without an external moral law there can be no genuine responsibility. Sartre's reply is that responsibility is internal to freedom itself and becomes visible precisely in concrete situations (Sartre, 2007; Sartre, 1992).

Methodologically, existentialists turn away from deductive rationalism and instead employ phenomenological and hermeneutic approaches. Influenced by Husserl's phenomenology and Heidegger's ontological hermeneutics, existentialist thinkers seek to describe existence as it is given in experience rather than as it is captured in conceptual systems. This shift reflects a deeper conviction: that the human being cannot be understood as an object among objects but must be grasped from the standpoint of lived experience (Guignon, 2004; Cooper, 1999). Existential analysis thus takes the

form of describing moods, choices, projects, and relations with others, all of which reveal the structure of existence from within. Taken together, these commitments reveal existentialism as a philosophy of protest against reductionist accounts of human life. It challenges the tendency to objectify being, whether through metaphysical ontology, scientific positivism, or universalist morality, and insists instead on grounding philosophy in the situated, finite, and often anguished condition of human existence. At the same time, existentialism is not merely descriptive but normative: it demands that human beings take responsibility for their freedom and confront the anxiety, ambiguity, and uncertainty that accompany it (West, 2010). For Sartre, this methodological commitment is decisive: the meaning of freedom and responsibility cannot be established by abstract deduction but must be shown through analyses of concrete experience – analyses that culminate in his phenomenological ontology in *Being and Nothingness* (Sartre, 1992).

From Sartre's perspective, grasping the principles and arguments of existentialism requires careful attention to his responses to the criticisms commonly directed against it. In *Existentialism Is a Humanism*, Sartre provides one of the clearest expositions of what existentialism means. He argues against charges of quietism, subjectivism, and moral relativism by showing that existentialism, far from being a philosophy of despair, is a philosophy that affirms freedom, responsibility, and human dignity (Sartre, 2007). However, Sartre's defensive clarifications in this lecture presuppose the more systematic ontological analyses developed in *Being and Nothingness*, where the structures of consciousness, negation, and being-for-others provide the conceptual basis for his arguments about freedom, responsibility, and action (Sartre, 1992; Crowell, 2012). This point prepares the way for understanding the distinctive elements of Sartre's thought, which both align with and depart from the broader existentialist tradition.

4. Sartre and existentialism

In *Existentialism Is a Humanism*, Sartre begins by setting out what existentialism is not, responding first to Marxist critiques. Communists accused existentialism of generating despair and restricting the possibility of action, while Marxist critics more broadly claimed that it emphasized the darker sides of human nature and thereby denied the hopeful aspects of life (Sartre, 2007, pp. 17–18; Flynn, 2006). This Marxist objection is especially significant because Sartre's own project is not limited to an abstract account of individual freedom; it explicitly raises the question of how freedom is lived and assumed within concrete historical and social situations – precisely the terrain on which Sartre later seeks a closer dialogue with Marxism (Flynn, 2006). These criticisms are not merely rhetorical; they directly concern the practical implications of Sartre's account of freedom, namely whether freedom culminates in political paralysis or, on the contrary, compels praxis and engagement. Alongside these criticisms, Sartre also takes aim at Catholic thinkers, who accused existentialists of being ignorant of reality and of abolishing values, and who further insisted that existentialism lacked any genuinely humanistic dimension (Crowell, 2012).

In response to these criticisms, Sartre begins by defining the very concept of existentialism: "Let us begin by saying that what we mean by 'existentialism' is a doctrine that makes human life possible and also affirms that every truth and every action implies an environment and a human subjectivity" (Sartre, 2007, p. 19). He goes on to distinguish between two types of existentialism, Christian and atheist, arguing that their most important shared characteristic is the principle that "existence precedes essence" (Sartre, 2007, p. 20). For Sartre, himself an atheist existentialist, understanding what existentialism means requires first clarifying what it means to say that existence precedes essence. However, the conceptual weight of this principle becomes fully intelligible only when it is situated within Sartre's phenomenological ontology developed in *Being and Nothingness*, rather than being treated solely as a programmatic slogan from a public lecture (Sartre, 1992).

4.1. Two interpretations of Sartre: despair vs. hope and responsibility

A recurrent debate in the secondary literature concerns whether Sartre's existentialism should be read primarily as a philosophy of despair or, on the contrary, as a philosophy of hope grounded in consciously assumed responsibility. The "despair" interpretation typically begins from Sartre's emphasis on abandonment: if God does not exist, there is no pre-given human essence, no external moral law, and no transcendent guarantee that can secure meaning in advance. Premises: (i) abandonment and the absence of objective, external values; (ii) the contingency of existence; (iii) the anxiety provoked by radical freedom. Conclusion: existentialism appears as a doctrine that isolates the individual, intensifies anguish, and risks collapsing into pessimism or political quietism. This line of reading often underlies the Marxist charge that existentialism restricts action by focusing excessively on inner experience and despair (Sartre, 2007; Flynn, 2006). By contrast, the "hope and responsibility" interpretation starts from the same premises but draws a different conclusion. Premises: (i) freedom is not an optional capacity but the ontological structure of human reality; (ii) to choose is unavoidable; (iii) each choice constitutes a commitment that implicates both the self and others. Conclusion: Sartrean freedom does not culminate in paralysis but in responsibility and praxis: meaning and value emerge through situated action rather than through external guarantees. On this reading, Sartre's talk of "despair" does not recommend passivity; it names the refusal to rely on illusions and the demand to assume one's situation lucidly, thereby making action possible and ethically weighty. Sartre's defense of existentialism as humanism thus functions as a practical clarification of a position whose conceptual basis is developed more systematically in *Being and Nothingness* (Sartre, 1992; Sartre, 2007).

To explain this, Sartre contrasts it with the opposite notion – that essence precedes existence – through a concrete example: "Consider a manufactured object, for example a book or a paper cutter. One knows that the paper cutter is made by a craftsman who has conceived of it and who has paid attention to the concept of a paper cutter and to a preexisting technique of production. In this way the paper cutter is defined by its essence prior to its existence" (Sartre, 2007, p.21). Sartre stresses here that for any manufactured object, its function, form, and purpose are determined before it

is actually brought into being; its essence therefore exists before its existence. The craftsman does not produce any object without first having defined its essence. This contrast is methodologically important: Sartre uses an ordinary example to show that "essence" names a prior blueprint only in the domain of artifacts, whereas human existence cannot be interpreted through such a blueprint without distorting what it is to be a human subject (Sartre, 2007; Sartre, 1992).

Human beings, by contrast, come into the world first – indeed, Sartre emphasizes that they enter the world in a condition of lack precisely because their essence does not precede their existence – and only afterward attempt to complete themselves. Unlike the craftsman's artifact, the human being is not designed in advance. In this framework, Sartre distinguishes between *being-in-itself* and *being-for-itself* (Sartre, 1992, pp. 100–102). Human beings are *beings-for-themselves*, while *being-in-itself* refers to things outside of humanity, that is, entities whose essence is predetermined. The human being, as a being-for-itself, brings forth its essence through existence and through its actions. By contrast, being-in-itself possesses a given nature, while being-for-itself – the human – has no fixed nature. Thus, for Sartre, every human being constructs its own nature only after existing, and always through its actions (Sartre, 1992, p. 110; Flynn, 2006, p. 45). In this sense, Sartre's doctrine is not a psychological claim about self-fashioning but an ontological claim about the structure of human reality: the for-itself is not a thing with a fixed "what," but a being whose "what" is perpetually constituted in and through praxis (Sartre, 1992).

For Sartre, who was deeply influenced by Husserl, there exists an intentional relation between consciousness and a given object. Consciousness is always a "consciousness of" something. For this reason, Sartre argues that the essence of consciousness emerges precisely through its intentional relation with the objects of the world, and he firmly rejects any attempt to reify consciousness. To claim that consciousness has an existence independent of objects would, for Sartre, contradict its very essence and amount to treating consciousness as another object (Sartre, 1992, pp. 11–12; Crowell, 2012). Here Sartre radicalizes Husserl's insight about intentionality: consciousness has no substantial "content" that could be isolated from the world, and this non-substantiality is precisely what opens the space for negation, distance, and freedom (Sartre, 1992; Crowell, 2012). Sartre's analyses of consciousness are further clarified through his distinction between *being-in-itself* and *being-for-itself*. Consciousness, which has no existence as an object in itself and never coincides fully with itself, belongs to the domain of being-for-itself. Being-for-itself is not a thing or an object; by contrast, things and objects, which lack consciousness, belong to the domain of being-in-itself. Thus, being-for-itself refers to the human being as a conscious existence. While being-in-itself coincides with its own being and thus possesses a fullness, being-for-itself exists as a lack, precisely because it is the being of consciousness.

The human being as a conscious existence is therefore a being that is not simply what it is, and is what it is not. In this sense, consciousness is always engaged in a process of transcending itself, of going beyond what it is in the direction of what it is not (Sartre, 1992, pp. 100–105). Hence Sartre's claim that "man is nothing other than what he makes of himself." For example, the essence of a table is determined in

advance – its design, materials, and function are defined before it comes into being. It exists in fullness as a being-in-itself. By contrast, the human being does not exist in such fullness; it is not predetermined in the manner of the table. What makes the human truly human is precisely this: its freedom. Sartre emphasizes that the human being cannot be defined once and for all but is always in the process of becoming. From the first moment of consciousness until death, the individual constantly recreates itself through its actions (Sartre, 1992, p. 110).

At the same time, Sartre maintains that human freedom emerges in relation to nothingness. For Sartre, there is no nothingness prior to consciousness. Nothingness comes into being only through consciousness; it cannot exist independently as a separate reality. As he states, "Nothingness is not; it is made-to-be; it does not nihilate itself, it is nihilated" (Sartre, 1992, p. 21). Since nothingness is "made-to-be" only through consciousness, it is the conscious being – the being-for-itself, that is, the human – that brings it forth. Because God does not exist, being precedes nothingness; nothingness arises only after being, and indeed through being itself (Flynn, 2006, p. 49). While nothingness derives its existence from consciousness, consciousness does not receive its being either from itself or from any external source. Nothingness therefore exists through being-for-itself, but being-for-itself does not require any external ground for its existence. In this respect, Sartre's account enters into a critical dialogue with Heidegger's analysis of the nothing: whereas Heidegger thematizes the nothing in relation to anxiety and the disclosure of Being, Sartre situates negation and nothingness within the immanent operations of consciousness itself – within the for-itself as the site of nihilation (Sartre, 1992; Heidegger, 1993; Flynn, 2006). Sartre's aim here is not to prove that God does not exist. Rather, his point is that even if God were to exist, nothing would fundamentally change: the human being, as a conscious being, is always a being-for-itself. Whether God exists or not, the human cannot possess a predetermined essence nor exist as something already defined. It is precisely for this reason that the human being, as the creator of nothingness, cannot resemble the being-in-itself, which has a fixed essence. This leads Sartre to one of his most fundamental conclusions: human beings are absolutely free. Freedom is not an attribute added to human existence but its very condition. For this reason, existentialist philosophy, from Sartre's perspective, must be understood as affirming that the human being is a project, a being that continuously designs itself through its choices and actions (Sartre, 2007, pp. 22–24). This "project" is not merely a personal plan; it names the fundamental way the for-itself temporalizes itself, relating to its future possibilities and thereby giving meaning to the present (Sartre, 1992).

At this point, Sartre's phenomenological analyses in *Being and Nothingness* make the practical force of his ontology visible at the level of concrete experience. A well-known example is the "keyhole" scene: a subject is absorbed in looking through a keyhole, but suddenly becomes aware of being seen by another. In that instant, the subject experiences shame – not as a moral feeling externally imposed, but as a disclosure of one's being-for-others. The subject recognizes itself as exposed, as an object for the Other, and thereby confronts the irreducibly social dimension of existence (Sartre, 1992). This shift from absorbed perception to reflexive self-awareness renders visible Sartre's claim that human reality is constituted through situations and

relations rather than through a fixed essence. Far from encouraging quietism, the scene shows that freedom immediately takes the form of responsibility: one must respond – by owning one's act, by fleeing, or by attempting to conceal one's freedom through bad faith. In this way, the keyhole analysis supports the present paper's central thesis that Sartrean existentialism is not a doctrine of despair but a philosophy of action grounded in freedom and responsibility. Seen in this light, the keyhole analysis is decisive: the same experience (shame before the Other) can motivate flight into bad faith (supporting the "despair" reading) or, conversely, can disclose responsibility in a concrete situation and thereby compel lucid commitment and action (supporting the "hope and responsibility" reading) (Sartre, 1992).

For Sartre, the human being, who is responsible for providing its own design, is a being of projects. Emerging as a being that projects itself into the future and becomes conscious of this projection, the human is not like a moss, a cauliflower, or a decayed object; rather, it is a subjectively lived project (Sartre, 1992, pp. 111–113). As is clear from this, the human being, in constituting its own project, must become what it designs itself to be. Because it is a conscious being, the human possesses the freedom of projection. It is precisely for this reason – that the human exists in a condition of lack – that it has the possibility of freedom. Where there is no lack, there can be no possibility of freedom. The condition of lack belongs only to the conscious being. Thus, the absence of a predetermined essence characterizes the human being. As a being of projects, the human exists by pursuing transcendent goals, and in doing so, sets aims for its praxis. These aims, in turn, give meaning to praxis itself. For Sartre, as a being of praxis, the human being attains a form of completion by attempting to overcome its own lack in praxis (Crowell, 2012).

A table, by contrast, cannot transcend itself nor move beyond its own being. Such transcendence belongs only to the reality of the human as a being marked by lack. This act, which condemns the human being to freedom, simultaneously provides the possibility of self-realization. It is this condition that also provides the possibility of rendering an absurd life meaningful. What makes life meaningful is the human reality itself. The human being, in accordance with its conscious existence and freedom, makes life meaningful through its actions (Sartre, 2007, p. 28). Accordingly, the practical question is not whether meaning exists in advance, but how meaning is constituted through situated choice, commitment, and responsibility (Sartre, 2007; Crowell, 2012).

In conclusion, the human being is a creature left entirely to itself. As Sartre puts it, the human is "nothing other than a project" (Sartre, 1992, p. 115). The human distinguishes itself from being-in-itself by designing its own essence through its actions. While an object has a fixed nature, the human being has no such nature. For this reason, the human is nothing other than its own project. It exists to the extent that it realizes itself and is nothing more than the sum of its actions. Therefore, the human reality cannot be separated from its freedom; it is only through the awareness of freedom – accompanied by the anxiety of existence – that the human realizes itself. Thus Sartre, the philosopher who condemned humanity to freedom, calls out as if directly to individuals: "You are free, therefore choose yourself; find your path!" (Sartre, 2007,

p. 30). What marks Sartre's distinctive position is his grounding of existentialism in the concept of project.

Sartre's view of the human being as a project necessarily condemns it to freedom. For Sartre, if God does not exist, then there is no pre-given design for humanity, and thus the human being must design itself through its own actions. "Man is condemned to be free, condemned because he did not create himself, yet nonetheless free, because once thrown into the world he is responsible for everything he does" (Sartre, 2007, p. 29). Freedom, therefore, is inseparable from responsibility: because the human being is free, it must also be held accountable for its choices. For this reason, Sartre argues that the existentialist does not believe in the irresistible power of passion, nor does he accept passion as a legitimate excuse for action. Rather, existentialism insists that every human being is responsible for his or her own passions and the actions that follow from them (Sartre, 1992, pp. 556–560). Moreover, Sartre emphasizes that the existentialist cannot rely on any external sign or divine guide to dictate what one must do. Each individual, left in a condition of abandonment, must decide alone. The human being, always in the process of constructing itself, does so only through action. Consequently, against the charge that existentialism leads to passivity, Sartre insists that it directs the human being toward action. Human existence is meaningful only insofar as it is enacted, and essence is defined only through deeds. One becomes what one does: cowardice or heroism is not predetermined, but chosen. As Sartre notes, "the coward makes himself cowardly, the hero makes himself heroic" (Sartre 2007: 35). What one is depends entirely on what one does, and at any moment the coward may cease to be cowardly, just as the hero may cease to be heroic. This is also the point at which Sartre's position intersects with the Marxist demand for praxis: although he rejects deterministic accounts of human action, his emphasis on situated agency and responsibility directly challenges the accusation that existentialism undermines collective struggle (Flynn, 2006; Sartre, 2007).

Thus existentialism cannot be called a pessimistic doctrine. On the contrary, it is deeply optimistic: by affirming that every human being writes his or her own destiny through action, it grounds philosophy in hope rather than despair (Flynn, 2006, p. 87). Human beings are not fixed, completed essences but unfinished beings, who must continually constitute themselves and, in doing so, choose an ethics. Since there is no God to serve as the creator of values, there can be no universal or pre-given morality; values must be created through human freedom and responsibility (Crowell, 2012). In this sense, Sartre's ethics is not a codified moral system but an ethical demand internal to freedom itself: to choose is to commit oneself, and such commitment inevitably implicates others and the shared human world (Sartre, 2007; Sartre, 1992). In this sense, Sartre identifies existentialism with humanism, for humanism is the doctrine that places the human being, rather than God, at the center of the universe. Accordingly, Sartre insists that his existentialism is nothing less than a form of humanism.

5. Conclusion

From the discussions above, it is evident that the most fundamental principle underlying Sartre's existentialism – its core arguments and orientations – is the idea that the human being is condemned to freedom. For the existentialist philosophers considered in the first part of this study, freedom appears as a possibility open to the individual, to be realized or rejected in accordance with one's choices. For Sartre, however, the human being is free whether or not it acknowledges or exercises this possibility. This is because, unlike being-in-itself, the human has no pre-given essence or completeness (Sartre, 1992, pp. 100–105). Accordingly, Sartre's thesis does not merely celebrate freedom as an ideal; it identifies freedom as an ontological condition of human reality that structures agency in every concrete situation.

The decisive reason why Sartre insists that the human is condemned to freedom lies in his account of nothingness. There is no nothingness prior to being; rather, nothingness comes into existence only through consciousness. In this way, consciousness, by "nihilating," generates the very condition that makes freedom possible. This perspective prevents any attempt to think of the human being apart from freedom. Sartre argues that human reality possesses the capacity to isolate itself through nothingness, and this capacity itself is freedom (Sartre, 1992, p. 21; Crowell 2012). Hence, since consciousness brings forth nothingness, the human being as a conscious existence constructs its own essence retrospectively through praxis. What constitutes the true structure of the human is therefore not merely its individuality but its social and practical dimension. As a being of praxis, the human necessarily transcends itself and exists in relation to others. This is precisely why Sartre's account is not a theory of resignation: "nihilation" does not lead to passivity; it names the distance from the given that makes choice, commitment, and responsibility possible (Sartre, 1992).

In responding to the criticisms leveled against existentialism, Sartre also emphasizes that it cannot be equated with materialism. For while materialism may allow us to think of the human as an object, human existence cannot be analyzed in the same way as a table. A table may be considered in isolation, but the human being exists only with and through others. For this reason, Sartre's subjectivity is not Cartesian: the cogito, for Sartre, also implies the recognition of others, for others are conditions of one's own existence (Sartre, 2007, p. 45; Flynn, 2006). Thus existentialism is, in the final analysis, a humanism. The 'keyhole' scene illustrates this point in concreto: the experience of being-seen discloses being-for-others and shows how responsibility emerges within an interpersonal situation rather than from an abstract moral law (Sartre, 1992).

Accordingly, Sartre's existentialism may be summarized in the following arguments: The human being

- can write its own destiny through action,
- must become conscious of its inescapable freedom,
- can overcome despair,
- is a being of project,

- is responsible for others, and
- creates values for itself, such that there can be no universal morality.

The fact that the human reality, which designs itself, possesses no essence prior to its existence entails the rejection of the very idea of a fixed human nature. Unlike being-in-itself, the human being has neither an essence nor a state of completeness. The human is a project, a being of design. For this reason, the condition of being human is identical with being free. As a conscious being, the human has the possibility of realizing itself, and this possibility necessarily imposes responsibility. Such responsibility does not lead to passivity but rather to action. To exist consciously within freedom is to live with responsibility, and this responsibility produces anguish (*angoisse*). For Sartre, freedom compels the human being to act, but it also makes the individual responsible not only for itself but also for others (Sartre, 1992, p. 553–560). This responsibility is closely related to the existential disclosures Sartre thematizes through anguish and nausea (*nausée*), which accompany the recognition of freedom and the contingency of the world.

Faced with this realization, some individuals attempt to escape freedom, refusing to take responsibility for their choices. Such a refusal produces what Sartre calls "bad faith" (*mauvaise foi*), the condition in which the human being denies its own freedom by reducing itself to the status of a thing, a being-in-itself (Sartre, 1992, pp. 86–89). The person in bad faith deceives himself, fleeing from responsibility and attempting to hide from the anxiety that accompanies freedom. Yet to be human is to be free, and to be free is to be responsible. Since essence follows existence, the human must act in the face of anguish; action is the only way to realize freedom. As Sartre insists: "It is not a screen which separates us from action, but it is rather the very condition of action; it is a part of action itself" (Sartre, 2007, p. 35). According to Sartre, the fundamental principles of existentialism can be summarized as follows:

- Existentialism rejects inaction and laziness; it calls the human being to act and take responsibility.
- It is an optimistic doctrine, affirming that meaning and value emerge through action.
- Existentialism is a form of humanism, placing the human being – not God – at the center of existence.
- God does not exist; therefore, the human being must design itself through its own choices.
- There is no universal morality; moral values are created through individual freedom and responsibility.

Sartre's distinctive contribution lies in grounding his existentialist arguments in a phenomenological ontology of the for-itself, nothingness, and projects, thereby transforming freedom from a mere possibility into an inescapable condition (Sartre, 1992; Crowell, 2012). In this context, Sartre holds that another principle of existentialism is the acknowledgment that the human being is necessarily free. This acknowledgment, in turn, entails principles that reject God, affirm action, promote optimism, ground

humanism, and ultimately deny the validity of a universal moral law. For Sartre, the essential content of human nature is that it remains an as-yet undetermined project. From this perspective, existentialism – when viewed through Sartre's lens – renders the human being necessarily free, imposes responsibility, and thereby offers not a philosophy of despair but the possibility of revolt. Accordingly, the practical implications of Sartre's existentialism may be summarized as follows:

The human being

- must avoid inaction,
- must be aware that it gives direction to its life,
- must embrace rather than deny the possibility of freedom,
- must render life meaningful through its actions,
- must assume responsibility rather than deceive itself, and
- must remain hopeful.

In conclusion, while all existentialists address certain common principles, Sartre's understanding of existentialism clearly proceeds along a distinctive trajectory. Above all, his analysis of the concept of freedom – though often the target of criticism – transforms freedom from a mere possibility into an inescapable condition, thereby reconfiguring the fundamental principles, arguments, and implications of existentialist philosophy.

The analyses presented here demonstrate that Sartre's existentialism cannot be reduced merely to the commonly shared axiom that "existence precedes essence." His distinctiveness lies in grounding this principle in the claim that the human being is necessarily free, and in interpreting freedom in terms of responsibility, action, and hope. From this perspective, Sartre's existentialism does not lead to despair or passivity, as is often suggested in the literature, but rather calls the human being to write its own destiny through action, to construct its essence through praxis, and to assume responsibility in relation to others. In highlighting these dimensions, this study seeks to underscore Sartre's divergence from his contemporaries and to contribute to the reinterpretation of existentialism as a philosophy of hope and action. Accordingly, Sartre's phenomenological descriptions and ontological arguments converge on a single practical implication: freedom demands responsibility and is realized only in action.

References

Blacham, H. J. (1967). *Six existentialist thinkers*. New York: Harper & Row.

Camus, A. (1991). *The myth of Sisyphus* (J. O'Brien, Trans.). New York: Vintage International.

Cooper, D. E. (1999). *Existentialism: A reconstruction* (2nd ed.). Oxford: Blackwell.

Crowell, S. (2012). *Existentialism*. Oxford: Oxford University Press.

Çüçen, A. K. (2018). *İnsan hakları* (3rd ed.). Ankara: Sentez Yayıncılık.

de Beauvoir, S. (1948). *The ethics of ambiguity* (B. Frechtman, Trans.). New York: Philosophical Library.

de Beauvoir, S. (1949). *The second sex* (H. M. Parshley, Trans.). New York: Vintage Books.

Flynn, T. (2006). *Existentialism: A very short introduction*. Oxford: Oxford University Press.

Guignon, C. (2004). *The existentialists: Critical essays on Kierkegaard, Nietzsche, Heidegger, and Sartre*. Lanham, MD: Rowman & Littlefield.

Heidegger, M. (1962). *Being and time* (J. Macquarrie & E. Robinson, Trans.). New York: Harper & Row.

Heidegger, M. (1993). *What is metaphysics?* (D. F. Krell, Trans.). In D. F. Krell (Ed.), Basic writings. New York: HarperCollins.

Jaspers, K. (1957). *Way to wisdom: An introduction to philosophy* (R. Manheim, Trans.). New Haven: Yale University Press.

Löwith, K. (1964). *From Hegel to Nietzsche: The revolution in nineteenth-century thought*. New York: Holt, Rinehart and Winston.

Macquarrie, J. (1972). *Existentialism: An introduction, guide and assessment*. New York: Penguin Books.

Marcel, G. (1952). *The Metaphysical Journal* (B. Wall, Trans.). Chicago: H. Regnery.

Marx, K. (1976). *Capital: A critique of political economy* (Vol. 1) (B. Fowkes, Trans.). Penguin Books.

Marx, K. (2000). *Economic and philosophic manuscripts of 1844*. In D. McLellan (Ed.), *Karl Marx: Selected writings* (2nd ed.). Oxford University Press.

Merleau-Ponty, M. (1964a). *The primacy of perception and other essays* (J. M. Edie, Trans.). Evanston, IL: Northwestern University Press.

Merleau-Ponty, M. (1964b). Eye and mind (C. Dallery, Trans.). In J. M. Edie (Ed.), *The primacy of perception* (pp. 159–190). Evanston, IL: Northwestern University Press.

Moran, D. (2000). *Introduction to phenomenology*. London: Routledge.

Nietzsche, F. (1969). *Thus spoke Zarathustra* (R. J. Hollingdale, Trans.). London: Penguin Classics.

Sartre, J.-P. (1992). *Being and nothingness: A phenomenological essay on ontology* (H. E. Barnes, Trans.). New York: Washington Square Press. (Original work published 1943)

Sartre, J.-P. (2007). *Existentialism is a humanism* (C. Macomber, Trans.). New Haven: Yale University Press.

Şerifoğulları, G. (2023). *Naşıl yaşamalıyız? Camus felsefesinin sinemayla serimlenişi*. Bursa: Sentez Yayıncılık.

Taylor, C. (1989). *Sources of the self: The making of the modern identity*. Cambridge, MA: Harvard University Press.

West, D. (2010). *Continental philosophy: An introduction* (2nd ed.). Cambridge: Polity Press.

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 328 – 345

SELF-REALIZATION AS A WAY TO FREEDOM: PLATO AND SPINOZA IN DIALOGUE

Junshuo Bian*

Abstract: What does true freedom mean? For Plato and Spinoza, freedom requires finding the true self and becoming like God. This paper demonstrates their philosophical communicability by showing that both thinkers ground freedom in a metaphysical division between two orders of existence, one external and temporally determined, the other internal and eternal, and that the ethical life consists in one's progressive identification with the latter. For Plato this means the soul's ascent toward the intelligible realm through self-mastery, requiring detachment from and devaluation of the sensible world. For Spinoza, freedom is not a transition between realms but a transformation in causation, from external determination to joining God's self-causation as *causa sui*. Since Extension and Thought are equal attributes of God, Spinoza does not devalue the sensible world, and self-realization is equally expressed in political life through the rational preservation of *conatus*. The comparison illuminates how Spinoza transforms the Platonic ideal by grounding freedom in a metaphysics that gives equal weight to the bodily and political dimensions of human life.

Keywords: Plato, Plotinus, Spinoza, freedom, self-realization.

1. Introduction

Beginning with God and ending with human freedom, Spinoza's *Ethics* aims at constructing a metaphysics that can guide human beings towards freedom. He defines freedom as self-determination, and says that a "thing is called free which exists from the necessity of its nature alone, and is determined to act by itself alone" (E1def7).[1] This conception of human freedom as self-determination has been brought into fruitful comparative dialogues with many philosophers in the academia, for example, Matthew Kisner links this concept with the Kantian concept of autonomy (*Autonomie*), finding in both thinkers a shared commitment to rational self-governance (Kisner, 2011, pp. 1–16). Genevieve Lloyd and Moira Gatens link it with the Stoic ideal of freedom, focusing on how both Spinoza and the Stoics locate freedom in the knowledge of necessity and the transition from passivity to activity (Lloyd & Gatens, 1999, pp. 41–57). In this paper, I propose that the Platonic tradition offers another meaningful point of comparison. The legitimacy of this comparative study faces a challenge in the first place, since Spinoza is explicitly opposed to Plato in a letter to Boxel, where he says "to me the authority of Plato, Aristotle, and Socrates is not worth much" (*Ep.* 56). Are there then any reasonable grounds to compare Platonism with Spinoza?

[1] All references to Spinoza's works follow the standard academic citation format and refer to Curley's translation.

*Universität Leipzig, Germany; ORCID: 0009-0000-4187-1257; bianjunshuo@gmail.com

©2026 Junshuo Bian. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

The following reasons support such a comparative approach. Firstly, Spinoza was deeply shaped by traditions that are themselves Platonized, for example, Christian scholasticism, Islamic medieval philosophy, Renaissance thought, Dutch Cartesianism, and the Kabbalah (Clark, 1959, pp. 273-276; Aanen, 2016, pp. 279-299; van Ruler, 2008, 159-175).[2] Secondly, and more decisively, Spinoza's critical distance from a thinker does not preclude substantive philosophical affinity with him. Spinoza reconstructs and criticizes Descartes throughout his work,[3] yet still inherits from him the central vocabulary of substance, attributes, and extension (Wagner, 2023, p. 14). The same logic applies here between Plato and Spinoza. Spinoza might not agree with certain Platonic doctrines, but he is indeed influenced by Plato's thoughts through mediating scholars that makes it meaningful to study these two together.

This comparative study also faces a second challenge. Do Platonism and Spinoza share any common ground in their conceptions of freedom?To answer this challenge, it is necessary to sketch the meaning of freedom in this paper through preliminary analysis of the texts from both philosophers.

The Greek equivalent for freedom is "ἐλευθερία", and in Plato's own texts, this word takes on different appearances.[4] One meaning of "ἐλευθερία" in Plato's texts is related to the legal status of free citizens in contrast to slaves. For example, in *Laws* he writes "if a slave kills a free man in self-defense, he shall be liable to the same laws as he that kills a father" (*Laws* 870d). Another derivative sense of this word is related to the independence of a *polis* from another *polis* (*Republic* 395c). The third sense concerns the qualities of a free person, which somehow is like a gentleman that has certain traits of liberal characters. For example, a free person moves his body in a certain way, goes hunting in a certain way, etc. (Laws 669c, 823e-824a). The fourth sense regards the leisure a free person has (*Theaetetus* 172c-d). The fifth sense is related to a democratic way of living, where people have freedom of speech (*Republic* 557b). The sixth sense concerns self-mastery or self-realization. It aims at acquiring justice of the soul. For example, Plato says that:

> If the better aspects of their minds win and steer them towards orderly conduct and philosophy, they live a wonderful, harmonious life here on earth, a life of self-control and restraint, since they have enslaved the part which allowed evil into the soul and freed the part which allowed goodness in (*Phaedrus* 256b).

The sixth sense of "ἐλευθερία" in Plato could be brought into dialogue with Spinoza since they share a common ground of self-mastery and self-determination. However, there is a third challenge for this comparative approach. As Hansen and Laks argue, this sense of the word remains relatively peripheral in Plato. Plato only sporadically associates the concept with self-mastery, which is more closely connected to his conception of justice (Hansen, 2010, p. 21; Laks, p. 133). It is thus a question of whether it

[2]Spinoza's library and biography also prove such an intellectual connection, see Wagner, *Spinoza's Neoplatonism*, 40–46.

[3]For example, *Ep.* 2, *Ep.* 30, *Ep.* 81, E5pref.

[4]There are different generalizations about the meanings of "ἐλευθερία" in the academia. See for example, Laks, "Freedom, Liberality, and Liberty in Plato's *Laws*", 130-152, Lane, "Placing Plato in the history of liberty", 702-718 and Stalley, "Plato's doctrine of freedom", 145-158.

makes sense to compare these two philosophers when the sixth sense is peripheral. I do admit that "ἐλευθερία" is sporadically linked with self-mastery, and that self-mastery is more related to justice. But there is a difference between saying that Plato rarely uses the word "ἐλευθερία" to describe self-mastery, and saying that Plato has no teaching about self-mastery that could be reconstructed as a teaching about freedom at all. Plato does, however, offer an ethical teaching centered on self-mastery and justice, one that is as central to his philosophy as the teaching of freedom is to Spinoza's. The aim of this paper, therefore, is to compare the conceptions of freedom in Platonism and Spinoza on the basis of their shared emphasis on self-mastery and self-realization, rather than on a strict correspondence of terminology.

With such a goal, I intend to show that for both Plato and Spinoza, freedom requires finding the true self and becoming like god. The possibility of realizing the true self lies in a metaphysical structure that offers two ways of living. I divide my reconstruction in the following parts: firstly I will argue that Plato and Spinoza share a comparable metaphysical framework based on a division between two worlds: one external, changing, and temporal, the other internal, unchanging, and eternal. Within their respective metaphysical frameworks, both Plato and Spinoza propose a particular way of life as the ethical goal of human beings. For Plato, this goal consists in attaining self-mastery and becoming like God, whereas for Spinoza it consists in attaining self-determination and participating in *causa sui*. I will then reconstruct their respective ethical teachings and bring them into dialogue with one another. This comparison is meaningful in two respects. First, it demonstrates the philosophical possibility of a dialogue between Platonism and Spinoza on the concept of freedom. Second, by bringing their respective views into dialogue, it allows for a deeper understanding of freedom as self-realization in both philosophers.

2. The division of two worlds

For a philosopher with a specific metaphysical framework, their ethics, if internally consistent, can be properly understood only in relation to their metaphysical conception of being. Certain metaphysical assumptions of a philosophy would entail a certain direction of how a person is supposed to live. This is true of both Plato and Spinoza. In Plato's philosophy, the ultimate moral goal of realizing the true self, which is justice, depends on a metaphysical division of two worlds. Consider these passages in the Republic about true philosophers:

> His eyes are turned to contemplate fixed and immutable realities, a realm where there is no injustice done or suffered, but all is reason and order, and which is the model which he imitates and to which he assimilates himself as far as he can. (*Republic* 500c)[5]
>
> For the gods will never neglect the man whose heart is set on justice and who is ready, by pursuing excellence, to become as like god as man is able. (*Republic* 613a-b)

[5]All references to Plato's works are cited by the standard Stephanus pagination.

This shows that freedom for Plato means to transcend the sensible world to the world of God. This is based on the metaphysical division between the world we experience and the world of God, which is the world of Forms. In Book VI of the *Republic*, Plato refers to the analogy of line to show the division between the two realms:

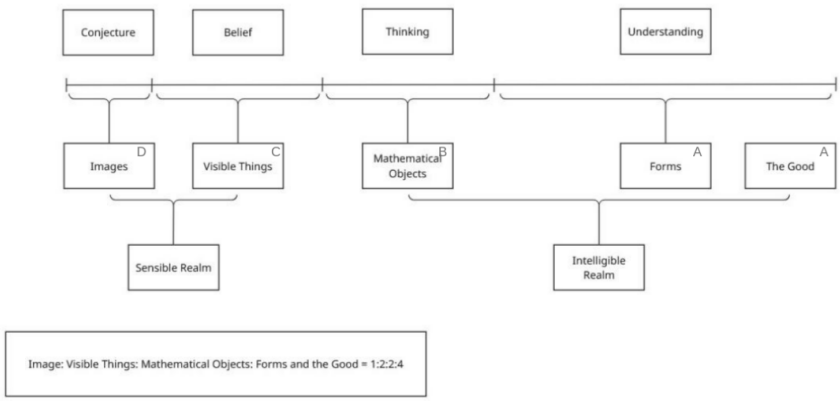


Figure 1. Division Between the Two Worlds.

Plato writes:

> Suppose you have a line divided into two unequal parts, and then divide the two parts again in the same ratio, to represent the visible and intelligible orders. This gives you, in terms of comparative clarity and obscurity, in the visible order one sub-section of images (D)[...] Let the other sub-section (C) stand for the objects which are the originals of the images[...]In one sub-section (B) the mind uses the originals of the visible order in their turn as images[...] in the other (A) it moves from assumption to a first principle which involves no assumption, without the images used in the other sub-section, but pursuing its inquiry solely by and through forms themselves. (*Republic* 509d-510b)

Plato's divided line analogy presents a carefully structured hierarchy that connects different levels of reality with corresponding modes of cognition. As Figure 1 shows, Plato distinguishes between two fundamental realms – the sensible and the intelligible – each characterized by distinct kinds of objects and corresponding modes of cognition. Within the sensible realm, cognition proceeds through the lower faculties of conjecture and belief. At its lowest level are images, such as shadows, reflections, and artistic representations. These occupy an ambiguous ontological position: they both are and are not what they represent, since their existence and identity depend entirely on the originals of which they are images. Above these images are the visible particulars – the physical objects, living beings, and artifacts of the sensible world that serve as originals for such images but are themselves only imperfect copies of higher realities. The intelligible realm, accessed through thinking and understanding, contains objects of greater reality and truth. The level of thinking is exemplified by

mathematical reasoning, where the soul employs hypotheses and visible diagrams to grasp abstract structures. Yet even these mathematical entities function as intellectual images of higher realities. At the summit of this hierarchy stand the Forms themselves, culminating in the Form of the Good, which transcends even Being and serves as the source of both existence and intelligibility.

The relationship between the sensible and intelligible realms presents one of the most challenging metaphysical problems in Platonic philosophy. The primary conceptual bridge Plato constructs between these realms is the notion of "participation" (μέθεξις) or "partaking", through which sensible particulars derive their characteristics from the transcendent Forms. However, to properly understand this relationship, we must first carefully examine the nature of the Forms themselves.

Following Harte's analysis, four key features characterize the metaphysical status and function of the Platonic Forms (Harte, 2011, pp. 193–194). First, the Forms possess ontological primacy: they constitute the most fundamental level of reality and embody true Being. Unlike sensible particulars, which are always in a state of becoming, the Forms remain identical with themselves and thus represent genuine, unchanging existence. They are "uniform and self-contained, remain always constant and invariable"(*Phaedo* 78d). Second, the Forms exhibit causal primacy and efficacy. Each Form serves as both the paradigmatic model and the explanatory cause for the sensible things that share in its character. For example, the Form of Beauty is not merely beautiful – it is Beauty itself, functioning as the source and explanation of beauty in all particular beautiful things. This means that "whatever else is beautiful apart from Beauty itself is beautiful because it partakes of that Beauty" (*Phaedo* 100c). Third, the Forms perform a semantic function, providing the metaphysical foundation that makes meaningful discourse possible. Plato says that "we always postulate in each case a single form for each set of particular things, to which we apply the same name" (*Republic* 596a). Finally, the Forms enjoy epistemic supremacy; Plato writes that "when the mind's eye is fixed on objects illuminated by truth and reality, it understands and knows them, and its possession of intelligence is evident; but when it is fixed on the twilight world of change and decay, it can only form opinions" (*Republic* 508d).

The range of entities that qualify as Forms is remarkably diverse. Different dialogues suggest different ranges of Forms, encompassing moral and aesthetic values such as Justice, Beauty, and the Good; quantitative properties such as Largeness and Equality; physical qualities such as Health and Strength; and even natural kinds and artifacts, such as the Form of the Couch. Some Forms occur in oppositional pairs, such as Largeness and Smallness, while others stand independently.[6] It is difficult to determine the precise range of Forms, but what is clear is that they function as the metaphysical grounds for the existence of sensible things.

The precise nature of the participation remains one of the most contested issues in Platonic scholarship, particularly regarding the question of whether Forms are

[6]As Annas notes, Plato offers no systematic, unified treatment of the Forms across the dialogues, and their scope varies considerably from text to text. See Annas, An Introduction to Plato's *Republic*, 217–218.

genuinely immanent in sensible things or maintain complete transcendent separation. Contemporary scholars have developed sophisticated and contrasting interpretations of this relationship. Some tend to argue that Forms are in most cases actually in the sensibles (for example, Harte, Fine), some tend to argue that Forms are completely separate (for example, Devereux). Fine offers a nuanced analysis of separation, proposing that a Form F is separate if and only if it can exist independently of any sensible particulars instantiating it (Fine, 2003, p. 263). Harte advances a critique of the two-world interpretation based on such an understanding of separation. Her argument draws particular force from a crucial passage in the *Phaedo* concerning the already mentioned forms of largeness and smallness:

> When you then say that Simmias is larger than Socrates but smaller than Phaedo, do you not mean that there is in Simmias both largeness and smallness? (*Phaedo* 102b3-5)

The interlocutor Echecrates agrees to the question, affirming that both qualities are "in" Simmias. For Harte, this passage suggests that the Forms – or at least their manifestations – are literally present within sensible particulars. Yet she recognizes an important ambiguity: what inheres in particulars may not be the Forms themselves but rather their immanent characters – the instantiated copies or images of the Forms that exist within particular things (Harte, 2011, p. 208).

Devereux, however, advances an argument for the complete separation of Forms from sensible particulars, grounding his interpretation in careful analysis of key passages from the *Phaedo* and *Timaeus* (Devereux, 1994, pp. 63–90). His reading hinges on a crucial distinction between Forms themselves and immanent characters.[7] The textual evidence for this distinction emerges most clearly in the mentioned *Phaedo*'s discussion of largeness and smallness. When examining how Simmias can be both larger than Socrates and smaller than Phaedo, Plato explicitly distinguishes between "largeness in us" and "largeness itself" (*Phaedo*, 103 b5). This terminological distinction, Devereux argues, reflects a fundamental ontological separation: while immanent characters (the "largeness in us") exist within and are dependent upon particular sensible objects, Forms themselves maintain complete transcendence and independence from the sensible realm.

Devereux finds definitive confirmation of this separatist interpretation in the *Timaeus*' characterization of Forms as:

> ...always the same, ungenerated and imperishable, neither receiving into itself anything from elsewhere nor itself going into something else in some place, invisible and imperceptible by any sense, it being that which is given to intellect to investigate. (*Timaeus* 51e–52b)

This passage, with its emphatic denial that Forms can "go into something else", appears to explicitly reject any notion of genuine immanence. According to Devereux,

[7] For more knowledge on immanent characters, refer to for example Fine, "Separation", 31–87.

this represents Plato's mature position from the *Phaedo* onward[8]: Forms maintain complete ontological independence from the sensible realm, related to particulars only through the mediation of immanent characters.

For me, the interpretive tension regarding the relationship between Forms and sensible particulars reflects a deeper complexity in Platonic metaphysics that resists overly simplistic resolution. While passages like *Phaedo* 100 d4–e3 clearly speak of the presence of Forms in particulars, equally compelling textual evidence mentioned above supports the Forms' transcendence and separation from the sensible realm. Thus, instead of attempting to settle this tension once and for all, we might note that much of the Platonic tradition has been interpreted as giving prominence to the transcendent dimension of the Forms and their distance from the sensible world. Such an emphasis appears to have shaped the historical unfolding of Platonism and its influence on later philosophical reflection; I prove this by referring to Plotinus.[9]

In Plotinus' philosophical system, the emphasis on transcendence reaches its clearer articulation through his doctrine of the three hypostases: the One, Intellect, and Soul. While all three hypostases transcend the material world, they exhibit different degrees and modes of transcendence. The One represents absolute transcendence – a principle beyond being itself that defies all predication and description. As pure unlimited unity, it stands as the ineffable source of all reality while remaining radically separate from everything it generates.

Below the One, both Intellect and Soul belong to the intelligible realm and contain the Forms. The creation of the material universe, as Plotinus conceives it, occurs primarily through the agency of the Soul. The generation of the sensible world occurs through Soul's inclination toward autonomy and self-expression, resulting in the production of sensible matter:

> It [Soul] generates something completely different from Itself …what is generated is actually no longer a form of soul – for it is no longer alive – but is total indefiniteness. (Enn. 3.4.1.5-12)

This "total indefiniteness" – sensible matter – functions as pure privation, serving as a receptacle for the Soul's creative activity. Plotinus employs the potent metaphor of a mirror to describe how matter reflects the intelligible Forms, producing mere images or imitations of true reality. The resulting sensible objects are thus "decorated corpses" – lifeless reflections that never genuinely participate in the Forms they appear to instantiate. Matter itself remains fundamentally unaffected, serving only as a passive medium for displaying distorted images of intelligible reality (Enn. 3.6.13.25-35).

[8]The stylistic analysis divides Plato's dialogues into three periods (early, middle and late); it is believed that *Phaedo* belongs to the middle period, and *Timaeus* belongs to the late period. For more knowledge of the intricacy of this division, see Leonard Brandwood, *The Chronology of Plato's Dialogues*.

[9]I conceive Platonism not merely as the philosophical system articulated by Plato himself, but as a rich and enduring intellectual tradition that encompasses subsequent thinkers who developed and transformed Platonic insights in significant ways. Among these later Platonists, Plotinus stands as perhaps the most profound and systematic expositor of Platonic thought, whose philosophical contributions warrant particular attention in our analysis. For discussion of this topic, see Gatti, "Plotinus", 10-37. All references to Plotinus' works are cited by the standard Enneads numbering (tractate, chapter, line).

This radical devaluation of the sensible world represents an intensification of Platonic transcendence. In this sense, the separation between intelligible and sensible reality provides the ontological foundation for the goal of Platonism's ethical imperative: the goal of "becoming like god". The very structure of reality, as conceived by Platonic thought, demands that genuine human fulfillment to realize themselves must involve transcendence of the sensible realm and ascent toward purely intelligible being. Since true reality resides exclusively in the transcendent realm of Forms, the philosophical life necessarily entails a progressive detachment from sensible concerns and an orientation toward intelligible truth.

Spinoza also has a metaphysical two world division, indicating two ways of existence, one temporal, the other eternal. His worldview is clearly structured: there is only one Substance, God, and this Substance is in itself and is conceived by itself. God is *causa sui*, the cause of itself, "whose essence involves existence, or that whose nature cannot be conceived except as existing" (E1def2). As *causa sui*, God follows only his own nature and is not forced by any external thing, being free in always affirming its essence as existence.

Everything other than God is called mode, which is in God and conceived through God. God expresses himself in infinitely many modes through infinitely many attributes, which include Thought and Extension. Modes include finite modes and infinite modes. A finite mode is for example a table, a human being, or an idea of an apple. An infinite mode is what comes from the absolute nature of God, which is some universal rule that governs every finite mode in a certain attribute. There are the immediate infinite mode and mediate infinite mode, the latter of which comes from the first. For example, the law of motion and rest is an immediate infinite mode of the attribute of Extension, and the face of the universe is a mediate infinite mode that comes from the law of motion and rest (*Ep. 65*). There are two ways to conceive the existence of a finite mode – one horizontal, the other vertical.[10] In the temporal domain, finite modes determine each other through a horizontal infinite causal chain between themselves. This causal chain is in endless self-movement, and thus every mode is inter-connected in the sense that a finite mode is constantly shaped and reshaped by this inner movement. To look into this process more clearly, I analyse the finite modes in the attribute of Extension, which are extended finite things, since according to E2p7, finite modes expressed under all the other attributes have the same order and connection with each other as extended finite modes. Spinoza says in E1p28:

> Every singular thing, or any thing which is finite and has a determinate existence, can neither exist nor be determined to produce an effect unless it is determined to exist and produce an effect by another cause, which is also finite and has a determinate existence; and again, this cause can neither exist nor be determined to produce an effect unless it is determined to exist and produce an effect by another, which is also finite and has a determinate existence, and so on, to infinity.

[10]This term is borrowed from Yovel, see Yovel "The Third Kind of Knowledge as Alternative Salvation", 157–175.

This means that finite modes limit each other to their finitude. Spinoza says in E1def2 that "that thing is said to be finite in its own kind that can be limited by another of the same nature". This definition states that between finite modes of the same nature, for example, bodies in Extension or ideas in Thought, there is a mutual limitation by each other, causing their finitude.

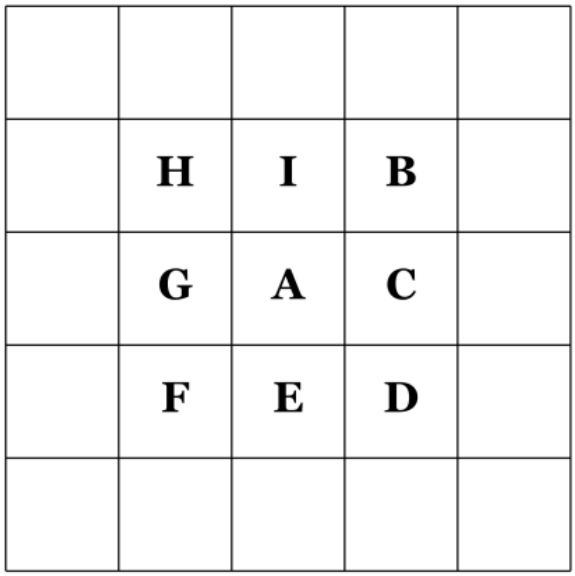


Figure 2. The horizontal causation.

Consider Figure 2. It depicts an interconnected world consisting of an infinite number of individuals under the attribute of Extension – A, B, C, D, E …I, and so on. There is no "truce area" between these modes in which a mode remains isolated from all others. This follows from Spinoza's rejection of a vacuum in the world of extended things (E1p15s): every surface of every finite mode borders other finite modes, whether larger or smaller. The finite mode A is then determined, or limited by B, C, D, E …I, which directly border its surface. These finite modes are in turn with the same reason determined, or limited, by all the finite modes that directly border their surfaces, *ad infinitum*. The finite mode A is therefore immediately limited by B, C, D, E …I and mediately limited by the finite modes limiting B, C, D, E …I, *ad infinitum*. Same goes for finite modes B, C, D, E …I in this horizontal picture.

The vertical causation of a mode concerns the eternal domain. Since God is immanent in all things, the ways that all things come about are determined by the necessity of God's nature. Thus Spinoza writes in E1p29, "all things have been determined from the necessity of the divine nature to exist and produce an effect in a certain way". What determines the existence of a finite mode in this manner is formal essence. But

what exactly is a formal essence?[11] When talking about the third kind of knowledge, Spinoza says in E2p40s2 that this kind of knowledge "proceeds from an adequate idea of the formal essence of certain attributes of God to the adequate knowledge of the formal essences of things." This seems to indicate that both the attributes of God and things have formal essences, and the ideas of these formal essences are all adequate. An adequate idea, according to Spinoza, has all the intrinsic denominations of a true idea. An adequate objective idea of a certain attribute or a certain thing, in this sense, defines the attribute and the thing. Thus, a formal essence is that which, when given, entails that an attribute or a thing is posited, and which, when taken away, entails that the attribute or thing is likewise taken away. We can therefore see, first, that formal essence is what defines an attribute of God or a singular thing in itself, independently of whether it is thought. A formal essence is what follows from God's infinite nature (E2p7c), and "all the things which follow from the absolute nature of any of God's attributes have always had to exist and be infinite, or are, through the same attribute, eternal and infinite" (E1p21). This means that formal essences are what eternally determine the existence of finite modes. Figure 3 shows this vertical causation:

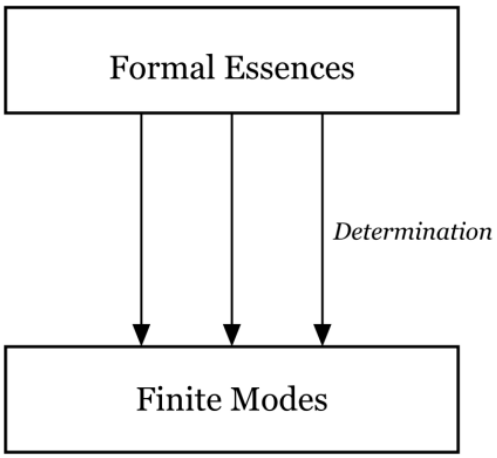


Figure 3. The vertical causation.

To understand the two ways of existence in Spinoza's philosophy, consider an example from E1p17s. Suppose there are two people – one person could be the temporal cause of the existence of the other if one is the other's parent. But this parent cannot be the eternal cause of the existence of the other, since this is determined by formal essence. The formal essence is eternally so, even if one of these two people does not exist. But if the formal essence does not exist, both of these two people would not exist. Thus, the same mode exists in two different senses: in the horizontal or temporal domain, it is determined by the infinite causal chain through time, while

[11]The concept of formal essence is a difficult topic, scholars debate over whether formal essences are singular or universal, and whether formal essences are infinite modes. For a basic understanding in this debate, see Lærke, "Aspects of Spinoza's Theory of Essence", 11–44.

in the vertical or eternal domain, it is determined through formal essence in God's infinite nature.

For Spinoza eternity and time are two completely different domains. Spinoza says that:

> For such existence, like the essence of a thing, is conceived as an eternal truth, and on that account cannot be explained by duration or time, even if the duration is conceived to be without beginning or end. (E1def8exp)
>
> The difference between Eternity and Duration arises from this. For it is only of Modes that we can explain the existence by Duration. But [we can explain the existence] of Substance by Eternity, i.e., the infinite enjoyment of existing, or of being. (*Ep.* 12)

It is clear that the ontological status of eternity differs categorically from that of duration, as substance or God, comprising infinite attributes and formal essences, exists as eternal truth. This eternal existence remains irreducible to and inexplicable through temporal phenomena characteristic of finite modes. Such a metaphysical assumption provides two ways of living for Spinoza – one externally determined, the other internally determined – to reach freedom as self-determination one needs to exist internally in eternity.[12]

3. Becoming like God as reaching the true self

Plato's and Spinoza's respective metaphysical divisions between two worlds suggest a path toward achieving genuine self-mastery and self-determination through the imitation of God.In Plato's case, it is with the soul that human beings can escape from the sensible world and become like God. Plato discusses the soul in many different places (for example, *Republic*, *Phaedo*, *Protagoras*, *Gorgias*, *Timaeus*), and he seems to have different accounts of whether the soul is composite (*Timaeus* 69e–72b) or incomposite (*Phaedo* 78b–80b). I analyse an account in *Republic* Book IV, where Plato argues that, just as a just city requires proper ordering of its classes (rulers, auxiliaries, producers), the soul achieves justice through the right relationship between its parts: reason, spirit, and appetite (*Republic* 434d–444e). Reason is a part that seeks truth and justice – it is considered as the true self. Spirit is the source of honor-loving desire, and is responsible for feelings of indignation. Appetite represents desires for all the sensible satisfactions, which stands for the untrue way of living.

These elements naturally experience tension–reason may counsel against what appetite desires, while spirit can either reinforce reason's authority or satisfy appetite's demands. For example, a person might have the desire to drink alcohol, but reason could tell her that it is not good at that moment to drink it, and after drinking it and feeling uncomfortable, the spirit in her soul functions as anger against herself. A

[12]Several scholars have mounted challenges to the strict distinction between temporality and eternity in Spinoza, suggesting eternity entails endless time. However, since time is considered by Spinoza as a mental construction lacking real existence (*Ep.* 12), it is hard to believe eternity equals endless time. See Donagan, "Spinoza's Proof of Immortality", 241-258, Matson, "Body Essence and Mind Eternity in Spinoza", 82-95, Kneale, "Eternity and Sempiternity", 223-238.

well-ordered soul requires reason to rule with wisdom while spirit acts as its ally in governing appetite. A well-functioning and noble soul should not have these clashes – reason should always maintain control of the situation and ensure that desire does not grow too much, with the help of spirit.

But being godlike requires more than just a well-ordered soul, since the divine realm remains unpolluted by sensible things. The physical world, with its constant becoming and change, presents only deceptive images of true Being in the intelligible realm through the unreliable senses. The soul's ultimate aim to control itself, therefore, is to escape this sensible world through separation from the body. Only then can it achieve pure self and, like the Forms and gods in the intelligible world, become truly unchanging and indestructible.[13]

How is this achieved? Plato offers a striking insight in the Phaedo:

> Ordinary people seem not to realize that those who really apply themselves in the right way to philosophy are directly and of their own accord preparing themselves for dying and death. (*Phaedo* 64a)

This appears to endorse suicide as realization of freedom, as the interlocutor in *Phaedo*, Simmias, immediately recognizes. He points out a tension: if death benefits philosophers by bringing them closer to their true Being, then suicide should be encouraged for reaching freedom.

Socrates responds by arguing that philosophers must wait for the proper time to die rather than hastening it themselves. His reasoning is theological: human beings are divine possessions under the gods' care. Just as masters would be angered if their property destroyed itself, the gods would be angered by unauthorized suicide.

As Bostock observes, this reasoning creates a significant tension. Death appears unambiguously good for the philosopher, yet divine command forbids it. The argument thus rests on authority rather than on a purely rational account of what is beneficial for freedom (Bostock, 2000, pp. 412–416). However, given that separation from body serves philosophers' interests, and their souls would achieve true self in the intelligible realm, it remains difficult to explain why they should delay this transition to care for mere images of true Being. Here I offer an argument for the importance of the earthly cultivation of justice which emerges from examining Plato's various accounts of reincarnation, which appear across multiple dialogues (*Gorgias*, *Republic*, *Phaedo*, *Phaedrus*, *Timaeus*). In these accounts, philosophical knowledge plays a crucial role in determining the soul's journey. The *Phaedrus* myth particularly illuminates this through its image of the soul as a charioteer with two horses – one well-behaved and one unruly. Initially, the soul travels like the gods along heaven's outer rim, dwelling in the realm of true Being and knowledge (*Phaedrus* 246a–249c). However, unlike the gods' perfectly controlled steeds, human souls struggle to maintain this elevation.

[13] It is to be noted that Plato has a different account regarding the immortality of the soul: sometimes he argues that this immortality is more like pure intellect, and sometimes he argues that the soul has memory and emotion. Tarrant contends with the help of *Apology* 41 a–c, where Socrates says that he could talk to the people from the old time, that the immortality of the soul is related to a complete soul. See Tarrant's introduction to *Phaedo*, 101.

Their unruly horses cause them to stumble and fall, losing their wings and entering the cycle of reincarnation into various lives – philosophers, artisans, businessmen, or even animals. This cycle typically lasts ten thousand years, with one crucial exception: philosophers who choose the philosophical life three consecutive times can return to the intelligible realm after just three thousand years.

This mythological framework suggests an argument against suicide's utility in achieving true self and freedom. Since returning to the intelligible realm requires sustained philosophical practice across multiple lives, premature death cannot accelerate this process. For non-philosophers especially, suicide would not hasten their return to divine contemplation. The path to godlikeness and true self requires the disciplined cultivation of philosophical wisdom rather than mere hastened escape from the sensible world. Thus, a person who is striving to reach their true self is supposed to cultivate justice, and this involves a well-ordered life that is not just personal, but also political. Following this line of interpretation, some scholars contend that freedom in Plato's philosophy requires imitating god politically. Analyzing key passages in the *Laws* (715e–718c), *Timaeus* (90b–d), *Alcibiades* (133b–c), and *Republic* (500c–d), Annas and Armstrong suggest that becoming like god means exercising reason in human affairs. On this view, the free human being imitates at the social level what god accomplishes at the cosmic level (Armstrong, 2004, pp. 171–183; Annas, 1999, pp. 52–71). These scholars refer to these two texts. In *Timaeus* 90b–d, Plato suggests that divinity is attainable through devotion to true wisdom and maintenance of proper psychic order under reason's guidance. Similarly, in *Republic* 500c–d Plato talks about how the philosopher, who gazes on the things that are always the same, tries to imitate the rational order of the universe by putting "what he sees there into effect in human behavior, both in private and public" (*Republic* 500d).

For Plotinus reaching the true self also hinges on "being god", but the path is rather unpolitical (Enn. 1.2.6.2-3). To achieve this end, one has to purify herself of the negative bodily feelings, and to drag herself up to the Intellect, which is her true self, being self-caused. In the process to become like god, one has to shake off all the earthly things, even the political virtues mentioned before. Plotinus writes:

> He will not locate the act of self-control in imposing a measure, but in separating him entirely as far as possible, absolutely not living the life of the good human being, which civic virtue values, but leaving this, and opting for another, the life of the gods. (Enn. 1.2.7.20-30)

This passage reveals how Plotinus radicalizes the otherworldly impulse already present in Plato's conception of freedom. Whereas Plato still maintained a political picture for self-realization, Plotinus places the true self solely in the intelligible realm. For him, genuine freedom begins when the soul withdraws from all external determinations, purifying itself of bodily affections and ascending toward the Intellect, which is its true self. In this ascent, the soul ceases to act as a participant within the ordered *polis* and instead becomes self-legislating, self-sufficient, and divine. Thus, Plotinus transforms the Platonic ideal of "becoming like god" into an ethics of absolute autonomy.

Reaching the true self also consists in becoming God, or *causa sui*, for Spinoza, because only God can be solely self-determined, and one can only join God to reach self-determination. Based on the division between the temporal and eternal domain, realizing the true self requires that one shares God's eternity as much as possible. Clyde Hardin advances a sophisticated interpretation of Spinoza's temporal theory, arguing that temporal order represents merely an "appearance" grounded in horizontal causal order, rather than an objective feature of reality. On this reading, adequate conception transcends temporal understanding, revealing instead the eternal logical entailments that connect an individual's *conatus*, or the actual existence of an individual, to formal essence and ultimately to God (Hardin, pp. 130–132). I agree with Hardin and find textual support for this distinction between temporal and eternal modes of existence in E5p29s:

> We conceive things as actual in two ways: either insofar as we conceive them to exist in relation to a certain time and place, or insofar as we conceive them to be contained in God and to follow from the necessity of the divine nature. But the things we conceive in this second way as true, or real, we conceive under a species of eternity, and to that extent they involve the eternal and infinite essence of God (as we have shown in IIP45 and P45S).

So a human being exists on two different levels; this suggests that genuine human freedom emerges through the transition from temporal to eternal modes of understanding and existence. This is realized by exercising adequate ideas. For Spinoza, there are three kinds of knowledge. The first one is imagination, which is from random experience and signs that are always "mutilated, confused, and without order for the intellect" (E2p40s2). The second one is reason, which concerns the common notions between things; it sees through the right order what equally is presented between things. The third one is intuitive knowledge, which grasps the formal essences of things in eternity. Imagination is the only source of error: it conceives everything through time, and cannot understand the real causal chain between things. The last two kinds bring adequate knowledge, since they offer true order of things beyond time.

To put this into an example, suppose a person who is driven by imagination, their behavior is always explained by external factors, which is considered as passive by Spinoza (E3def2). Sometimes they see something as popular, and their wish to keep up with the trend makes them buy certain popular items. Sometimes they like or dislike something based on others' opinions. They take their confused and inadequate desires to represent what they truly want, even though these desires are actually determined by the horizontal causal chain among finite modes. In this condition, a person is continually determined by external things rather than being self-determined. By contrast, through the other two kinds of knowledge, a person can recognize the eternal dimension of their existence and gain a clearer understanding of what constitutes their essence. On the basis of this understanding, their actions follow from their own essence and can be explained through their nature. Spinoza characterizes a person who acts in this way as active (E3def2), and it is in this sense that an active person can be understood as self-determined.

Hannah Laurens develops this theme through her conception of "twofold existence", arguing that "we are finite and our existence is in time, but simultaneously our essence instantiates timeless substance" (Laurens, p. 105). This duality manifests in two distinct modes of cognition and being: when operating through imagination, humans remain bound within duration, subject to the deterministic chain of finite modal causation. However, through the exercise of reason and intuitive knowledge, humans can achieve adequate understanding, participating in a changeless mode of existence that directly instantiates God's eternal essences.

In this eternal mode, human actions emerge not from external determination but as necessary expressions of internal essence. Laurens captures this in her formulation "finity in infinity" – the possibility for finite modes to participate in infinite, eternal existence through adequate understanding. This interpretation presents human freedom not as liberation from necessity but as the achievement of a form of necessity that flows from one's own nature as an expression of divine essence.

This understanding could be proven by Spinoza's texts about the intellectual love of God (*amor Dei intellectualis*). With adequate knowledge, a human being understands their essence in eternity. Such an understanding of eternity would bring the greatest satisfaction (E5p27), since "The Mind's intellectual Love of God is the very Love of God by which God loves himself, not insofar as he is infinite, but insofar as he can be explained by the human Mind's essence, considered under a species of eternity" (E5p36). This intellectual love of God participates in God's self-causation, since what is eternal is God (E5p30d). If one exists through one's own formal essence insofar as it is eternal, one exists in the manner of *causa sui*, determined solely by one's own nature. This self-causation transcends time, in the sense that a human being "cannot be absolutely destroyed with the Body, but something of it remains which is eternal" (E5p23).

However, as human beings also exist in the temporal domain – in this domain – one cannot fully get rid of the external determination and enjoy complete internal self-causation. The practical plan for a human being to become *causa sui* in this domain is to imitate *causa sui* politically. This imitation consists in preserving one's actual essence, the *conatus*, in the temporal domain through reason. If a human being, by reason's direction, knows how to preserve their finite self, they are acting from their self-preservation drive, and such a drive is one's nature (GDA 1). To act as much as possible from their own knowledge about what they need to benefit themselves requires collaboration with other human beings, whose nature is in agreement with each other (E4p35). In this way people cooperate in a political collectivity for their self-preservation, and this process is a process where one acts as much as possible from their own *conatus* drive. And even though one cannot exist temporally outside of external determination, such a rational collaboration would give one the most possible approximation of *causa sui* in the temporal domain.[14]

[14]Laurent Bove has a detailed understanding of this political project of imitating *causa sui*, see Bove, "Affirmation and Resistance in Spinoza", 8.

4. Way of self-realization: A dialogue between Plato and Spinoza

By understanding freedom as self-realization, I have found both in Plato and Spinoza an ethical project of self-realization based on their metaphysical assumptions. Several communicable points are to be noted. Firstly, while both philosophers understand freedom in the sense of self-realization, they have different focuses. For Plato, realizing the true self consists in moving from the sensible realm to the intelligible realm. The intelligible realm is a destination for human beings to reach, which concerns a re-ordering of the elements in the soul. Self-realization, then, is a process to organize in chaos a well-ordered soul. For Spinoza, however, self-realization relies more on different ways of causation. If a human being exists temporally, this person is externally determined by other finite modes in the horizontal causal chain. If the person exists eternally, this person, by acting solely from their own formal essence or nature, joins God's self-causation and is internally determined. This understanding of self-realization, instead of realizing a well-ordered soul, seeks to emphasize the identification between freedom and necessity. As Spinoza says in E1p29, "all things have been determined from the necessity of the divine nature to exist and produce an effect in a certain way" (E1p29). When one exists by being externally determined, one does not internalize the self-causation of God into themselves. Once this self-causation is internalized, one causes their own existence by joining God's necessity of nature.

Secondly, both philosophers maintain a differentiation between temporality and eternity, but the division of two worlds is constructed on a different basis. As Harry Wolfson articulated, the conceptual history of eternity in philosophical thought reveals a fundamental dichotomy: the Platonic conception, which posits eternity as fundamentally distinct from and irreducible to duration (*Timaeus* 37 C–D), and the Aristotelian view, which understands eternity as "sempiternity" – endless duration, exemplified in Aristotle's treatment of the universe and celestial spheres as eternal yet temporal entities (Wolfson, 1934, pp. 358–369). Spinoza takes on the Platonic concept of eternity; however, unlike Plato who links the sensible to time and devalues the sensible realm, Spinoza links external causation to time and maintains the equality between Extension and Thought. Extension and Thought are equal attributes of God, both constituting his infinite nature (E2p7s). In this sense, the world of bodies and the world of ideas are the same God expressed in parallel manners. Rather than understanding solely the extended world as temporal, Spinoza understands everything externally determined as temporal, which includes also thinking things. On an eternal level, both the formal essences of bodies and minds join the self-causation of God. To reach eternity, then, consists not in devaluating the sensible, but in realizing self-causation in both attributes.

This leads to the different attitudes of politics between Platonism and Spinoza. For Plato, reason as the true self exists in the intelligible realm, and the ultimate goal for human beings is to escape from the sensible world to the intelligible realm. Even though there exists an earthly practice for cultivating justice in the process of becoming like god, freedom as realizing the rational true self in the intelligible realm essentially does not value politics too much. This is evident both in Plato and Plotinus. Plato says that the philosopher king (the free human being) is active in politics not

because it benefits themselves, but because it benefits the citizens (*Republic* 347d). And as demonstrated in the last part, reaching the true self for Plotinus requires that one get rid of bodily influences, which renders politics rather meaningless for realizing freedom. For Spinoza, however, eternity does not entail a devaluation of the extended world, but rather a transcendence of temporality more generally. Consequently, what reason prescribes as the path toward eternity also concerns one's existence in the extended world. This is evident in Spinoza's political ideal, where he says that "the end of the Republic is really freedom" (*TTP* III/241). A free human being, in this sense, not only seeks being self-caused in mind, but also in body. To realize that in the bodily aspect, Spinoza gives a political picture where reason guides humans into affirming their self-preservation essence in a collectivity, because "the good which everyone who seeks virtue wants for himself, he also desires for other men; and this Desire is greater as his knowledge of God is greater" (E4p37).

As this paper has argued, for both philosophers, freedom understood as self-realization is not a matter of choosing among alternatives. Rather, freedom is one's progressive turning toward what is most real in themselves, and thereby toward what is eternal in the order of being. From this perspective, the relation between Plato and Spinoza becomes more evident. When Spinoza defines freedom as self-determination according to the necessity of one's own nature, he echoes the Platonic point of view whereby the true self is identified with the soul's ascent in the world of Forms. The ethical ideal of *amor Dei intellectualis* reproduces the Platonic ascent of the soul toward the intelligible and the good. The comparison between Plato and Spinoza on freedom thus proves philosophically fruitful, because placing them in dialogue reveals a shared philosophical grammar. This grammar, articulated in two different metaphysical languages, illuminates the ideal of freedom as self-realization across the history of philosophy.

References

Aanen, J. (2016). The Kabbalistic Sources of Spinoza. *Journal of Jewish Thought and Philosophy*, 24(2), 279–299.

Annas, J. (1999). Becoming Like God: Ethics, Human Nature, and the Divine. In J. Annas, *Platonic Ethics: Old and New* (pp. 52–71). Cornell University Press.

Armstrong, J. (2004). After the Ascent: Plato on Becoming Like God. In D. Sedley (Ed.), *Oxford Studies in Ancient Philosophy XXVI* (pp. 171–183). Oxford University Press.

Bostock, D. (2000). The Soul and Immortality in Plato's Phaedo. In G. Fine (Ed.), *Plato 2: Ethics, Politics, Religion, and the Soul* (pp. 404–424). Oxford University Press.

Bove, L. (2023). *Affirmation and Resistance in Spinoza: The Strategy of the Conatus* (H. Sharp & É. Filion-Donato, Trans. & Eds.). Edinburgh University Press.

Clark, M. (1959). How Plotinian is Spinoza's Doctrine of Freedom? *New Scholasticism*, 33(3), 273–290.

Devereux, D. (1994). Separation and Immanence in Plato's Theory of Forms. In C. C. W. Taylor (Ed.), *Oxford Studies in Ancient Philosophy XII* (pp. 63–90). Clarendon Press.

Fine, G. (2003). Separation. In G. Fine, *Plato on Knowledge and Forms* (pp. 263–287). Clarendon Press.

Gatti, M. L. (1999). Plotinus: The Platonic Tradition and the Foundation of Neoplatonism. In L. P. Gerson (Ed.), *The Cambridge Companion to Plotinus* (pp. 10–37). Cambridge University Press.

Hansen, M. H. (2010). Democratic Freedom and the Concept of Freedom in Plato and Aristotle. *Greek, Roman and Byzantine Studies*, 50(1), 1–27.

Hardin, C. L. (1977). Spinoza on Immortality and Time. *Southwestern Journal of Philosophy*, 8(3), 129–138.

Harte, V. (2011). Plato's Metaphysics. In G. Fine (Ed.), *The Oxford Handbook of Plato* (pp. 193–208). Oxford University Press.

Kisner, M. J. (2011). *Spinoza on Human Freedom: Reason, Autonomy and the Good Life*. Cambridge University Press.

Laks, A. (2007). Freedom, Liberality, and Liberty in Plato's *Laws*. *Social Philosophy and Policy*, 24(2), 130–152.

Laurens, H. (2012). Finite in Infinity: Spinoza's Conception of Human Freedom Explained Through His Metaphysics. *Stance: An International Undergraduate Philosophy Journal*, 5(1), 97–109.

Lærke, M. (2017). Aspects of Spinoza's Theory of Essence: Formal essence, non-existence, and two types of actuality. In M. Sinclair (Ed.), *The Actual and the Possible: Modality and Metaphysics in Modern Philosophy* (pp. 11–44). Oxford University Press.

Lloyd, G., & Gatens, M. (1999). *Collective Imaginings: Spinoza, Past and Present*. Routledge.

Plato. (1993). *Phaedo* (H. Tredennick & H. Tarrant, Trans.). Penguin Books.

Plato. (2000). *Timaeus* (D. J. Zeyl, Trans.). Hackett Publishing Company.

Plato. (2002). *Phaedrus* (R. Waterfield, Trans.). Oxford University Press.

Plato. (2003). *The Republic* (T. Griffith, Trans.; G. R. F. Ferrari, Ed.). Cambridge University Press.

Plato. (2015). *Theaetetus* (C. Rowe, Ed.). Cambridge University Press.

Plotinus. (2018). *The Enneads* (G. Boys-Stones, J. M. Dillon, et al., Trans.; L. P. Gerson, Ed.). Cambridge University Press.

Spinoza, B. (1985). *The Collected Works of Spinoza* (Vol. 1, E. Curley, Trans. & Ed.). Princeton University Press.

Spinoza, B. (2016). *The Collected Works of Spinoza* (Vol. 2, E. Curley, Trans. & Ed.). Princeton University Press.

Van Ruler, H. (2008). Substituting Aristotle: Platonic themes in Dutch Cartesianism. In D. Hedley & S. Hutton (Eds.), *Platonism at the Origins of Modernity: Studies on Platonism and Early Modern Philosophy* (pp. 159–175). Springer.

Wagner, J. (2023). *Spinoza's Neoplatonism: The Metaphysics of the Unitive Knowledge and Love of God* (Doctoral dissertation, University of Cambridge). Apollo Repository.

Wolfson, H. A. (1934). *The Philosophy of Spinoza* (Vol. 1). Harvard University Press.

Yovel, Y. (1990). The Third Kind of Knowledge as Alternative Salvation. In E. Curley (Ed.), *Spinoza: Issues and Directions: The Proceedings of the Chicago Spinoza Conference* (pp. 157–175). Brill.

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 346 – 350

Short communication

REASSESSING PLETHON'S CRITIQUE OF ARISTOTLE: RESPONSES FROM TRAPEZUNTIOS AND BESSARION

Márton Adorján Kecskés*

Abstract: Georgios Gemistos Plethon's critique of Aristotle, most fully articulated in *De differentiis Platonis et Aristotelis*, emerged from a context of intellectual and pedagogical struggle within late Byzantine philosophy. Advocating a revival of Platonism, Plethon framed Aristotle as a metaphysical and theological rival, emphasizing differences in accounts of form, ethics, and divine causation. This article argues that while Plethon identifies genuine tensions between Platonic and Aristotelian traditions, his presentation often simplifies Aristotle's positions and relies on mediated interpretations inherited from later scholastic and Arabic sources. In metaphysics and ethics alike, Plethon tends to sharpen contrasts to serve his polemical aims, sometimes at the expense of interpretive precision. Comparison with contemporaries such as Scholarios and Bessarion shows that alternative, more conciliatory readings were available. Ultimately, Plethon's contribution lies less in offering a balanced exegesis of Aristotle and more in reopening philosophical debate and redirecting attention toward Platonic thought in a predominantly Aristotelian intellectual climate.

Keywords: Plethon; Aristotle; Georgios Trapezuntios; Bessarion; Byzantine philosophy; Platonism.

1. Introduction

Georgios Gemistos Plethon (c. 1355 – 1452) sets out his fullest attack on Aristotle in De differentiis Platonis et Aristotelis, written during the years of the Council of Florence. He is not neutral – he pleads for a revived Platonism and seeks to displace the Aristotelian scholasticism that had come to dominate Byzantine instruction. That political and intellectual urgency shapes everything he says. Read as advocacy, Plethon's criticism identifies real tensions between Plato and Aristotle (Tatakis, 2003). Read as exegesis, it often falls short; and that gap between aim and method is important for assessing fairness.

While there is currently no single monograph dedicated exclusively to the full scope of Plethon's Aristotelian critique, the subject has been framed by several foundational studies. François Masai (1956) famously positioned Plethon's Platonism as a pagan challenge to Christian Aristotelianism, while C. M. Woodhouse (1986) provided the essential biographical and political context for these philosophical disputes. More recently, Vojtěch Hladký (2014) has offered a comprehensive analysis of Plethon's system, grounding De differentiis in its late Byzantine and ancient Neoplatonic roots. This

*National and Kapodistrian University, Athens, Greece; misteradyesl@gmail.com

©2026 Márton Adorján Kecskés. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

article builds upon these standard works by focusing specifically on the hermeneutic strategies used by Plethon and his immediate critics. This hermeneutic tension is most immediately visible in the realm of metaphysics, where Plethon's insistence on transcendental depth directly clashes with Aristotelian immanence.

2. Metaphysical critiques

Start with metaphysics. Plethon insists that Plato gives primacy to transcendent Forms; Aristotle, by contrast, locates form immanently in substances. That contrast matters. Late Platonists emphasize hierarchical levels of being (Hladký, 2014; Karamanolis, 2002); Aristotle insists that form and matter are inseparable in concrete things. Plethon sees the latter as a kind of reduction, a loss of metaphysical depth. There is something to that reading; yet Plethon tends to simplify Aristotle. He treats Aristotle as if the older philosopher simply denied the intelligible realm, whereas Aristotle gives a different account: forms remain central, but as formal causes within particulars. Moreover, Plethon's grasp of Aristotle is sometimes filtered through later medieval and Arabic commentaries; he reads Aristotle as if through a stack of interpretive intermediaries. That mediation explains some bluntness. (Steiris, 2017).

3. Ethical critiques and the doctrine of the mean

Ethical disagreement appears more pronounced at first glance. Plethon is more comfortable with moral ideals understood as stable and intelligible standards, which leads him to view Aristotle's doctrine of the mean with suspicion. In his moral critique Plethon addresses the doctrine of the mean (Aristotle), by construing the mean as simply a quantitative midpoint and drawing paradoxical conclusions about the construal, such as the possible "half-wicked" occupants of the mean (Plethon, 1973, pp. 328-31). Close analysis reveals that Plethon generalizes a single Aristotelian locution and makes it the center of the theory. According to the discussion of the *Nicomachean Ethics*, the mean discussed by Aristotle is determined in "relation to us" and is grounded in right reason and practical wisdom (*phronēsis*); the mean is ordered, not in an arithmetic sense, but in a teleologically and normative way (Aristotle, *Nicomachean Ethics* 1106b36-1107a2).

This presentation, however, oversimplifies Aristotle's position. For Aristotle, the mean is relative to us, but it is determined through practical wisdom (*phronēsis*) and deliberation rather than calculation (Adamson, 2022). Habituation and reflective judgment play a central role in the formation of virtue. The contrast Plethon identifies is therefore real, but his account does not fully capture the qualitative structure of Aristotelian ethical judgment.

4. Theological disagreements

Beyond the qualitative structures of virtue, the most significant doctrinal rift for later 15th-century readers concerned the ontological status of the divine. The theological disagreement is especially significant for later readers (Masai, 1956). Plethon

maintains that Plato's account of divine causation better accommodates a providential God, whereas Aristotle's unmoved mover cannot be described as creator in Christian terms. Aristotle does not posit a creator ex nihilo (Adamson, 2022); Aristotle's God functions principally as final cause, the object of love and intellect rather than an efficient maker who brings the world into being from nothing. In Aristotle's own idiom the divine is pure actualisation and the ultimate telos of motion and thought. Difficulties arise when Plethon treats this difference as evidence of theological deficiency. Aristotle's unmoved mover remains transcendent and ontologically distinct from the cosmos, even if it is articulated in non-Christian terms. Plethon thus intensifies a doctrinal contrast into a polemical judgment. While Plethon's polemic aimed to spark a Platonic revival, it effectively launched a multi-generational debate. The first major counter-offensive came from Georgios Trapezuntios, who sought to protect Aristotelian logic and natural science from what he saw as Platonic rhetorical excess.

5. Georgios Trapezuntios: Aristotelian method and polemic

According to the *Comparatio* of George of Trebizond, Aristotle is the greater tool of natural philosophy and logic, and many Platonic dialogues are rewritten as a political or a morally questionable thing to do (George of Trebizond, 1457). His approach is offensive: dialogical or pedagogical excerpts in Plato are interpreted as protest of intellectual hazard and rhetorical bombast. In his works, Monfasani demonstrates that the choices made by Trapezuntios are both very selective and rhetorically exaggerated; he is likely to transform the hermeneutic disagreements into the indictments of a culture (Monfasani, 1976, pp. 11–12).

The key point made by Trapezuntios is valid in a single register: the categories and the causal schemes provided by Aristotle are invaluable where the explanatory goal is empirical or scientific. But he often does go too far in the extension of a disciplinary benefit to a demand of philosophical totality – of treating a methodological pre-eminence in natural philosophy as an implied pre-eminence in the rest. In addition, his polemical craft tends to replace exegesis with invective, and in so doing, distorts dialogic genres and fails to exploit late antique interpretive resources that mediate between Plato and Aristotle (Monfasani, 1976, pp. 11-12).

A closer examination of the style of Trapezuntios reveals the price of his polemical style. In his criticism of Plato, when he interprets his mythic or dramatic instruments as philosophical fallacy, he frequently loses sight of the pedagogical end of his instruments and of the fact that Plato deliberately performs those instruments in order to generate inquiry and not to propound ultimate theses. The fusion of genre and doctrine brought about by Trapezuntios consequently also has a rhetorically destructive yet philosophically incomplete effect: it manages to dramatize the differences but does not attempt to extend to the inner logic of Platonic dialectic or even to the late antique exegetical traditions reinterpreting Plato on theological grounds (Monfasani, 1976, pp. 11–12). In this way the rhetorical vigor of Trapezuntios produces a significant defence of Aristotelian science, at the same time disabling the plausibility of his interpretive assertions.

6. Bessarion: contextualization, philology, two-register solution

Three methodological responses are offered by Bessarion in *In calumniatorem Platonis*: the precise fixation of the text, its reconstruction in context, and the expression of a different set of explanatory registers to Plato and Aristotle (Steiris, 2023). He shows that a lot of the accusations made by Trapezuntios are based on a mistranslation, his selective quotation, or even removal of passages to their argumentative context (Bessarion 1927/1469, Mohler 3.11.1.5–12). Where Trapezuntios interprets Plato as a failure of rhetoric as philosophy, Bessarion places Platonic modes in a theological and metaphysical program centered on the Good; he, in his turn, acknowledges the epistemic superiority of Aristotle in purely natural and analytical inquiry (Del Soldato, 2014, pp. 67–71).

Two close readings are an illustration of the way Bessarion does it. To start with, on metaphysical primacy Bessarion demonstrates that the supposed duality (Forms vs. particulars) can be in natural cases reduced to a difference of explanatory register: metaphysical enquiries about ultimate reality and providence are better approached with Platonic symbolism and procession, whereas empirical causation is better treated with Aristotelian four-cause analysis. The recovery of the broader textual contexts dismanages a number of the contradictions of which Trapezuntios takes advantage (Bessarion 1927/1469, Mohler 3.11.1.5–12). Second, regarding ethics, Bessarion shows that the mean of Aristotle does not represent a simple middle ground but involves practical wisdom; reductive accounts of Plethon do not actually participate in the pragmatics of the doctrine that render it normative (Fink, 2011, pp. 484–87). The interventions of Bessarion rely not only on the corrected readings, but also on the methodological economy: he subjects the disputed passages to philological examination and places them in the argumentative lines that display specific intentions. At the level of flattening genre in the rhetoric of Trapezuntios, and of abstracting forceful lines into the generalizations of the program in Plethon, a certain authorial context is restored in Bessarion, and a certain possibility of coexistence of various explanatory projects is demonstrated. With this differentiation between theological language (in which Platonic symbolism does conceptual labor) and empirical argument (in which Aristotelian causality gives meaning to mechanism), Bessarion provides a more secure platform on which to evaluate conflicts, as well as borrow from both traditions without reducing them to one another (Bessarion, 1927/1469; Del Soldato, 2014, pp. 67–71).

7. Conclusion

When viewed through the prism of these 15th-century reactions, Plethon's original critique appears less as a definitive exegesis and more as a catalyst for a broader intellectual realignment. From this perspective, Plethon's intervention does not seem to be a settled judgment, but rather as a provocation of sorts. His turn toward Plato mattered in a context where Aristotelian commentary dominated philosophical teaching, and it helped redirect attention toward Platonic texts and problems. At the same time, the manner in which Plethon argues often tends to sharpen contrasts in order to advance his own case of course. Aristotle is treated primarily as a rival position

rather than as a great thinker to be reconstructed on his own terms and merits, and this has consequences for interpretive accuracy. Plethon's role, then, I think, lies more so in reopening philosophical debate rather than in providing a reliable and accurate account of Aristotle's views.

Acknowledgments: An earlier version of this research was developed during a postgraduate seminar at the National and Kapodistrian University of Athens. The author is grateful to Professor Georgios Steiris for his scholarly guidance, the foundational bibliography, and his insightful feedback during the course.

References

Adamson, P. (2022). *Byzantine and Renaissance philosophy*. Oxford University Press.

Aristotle. (n.d.). *Nicomachean ethics*. (Referenced passages: 1106b36–1107a2.)

Bessarion, B. (1927). *In calumniatorem Platonis* (L. Mohler, Ed.). Schöningh. (Original work published 1469)

Del Soldato, E. (2014). *Against the calumniator of Plato (Bessarion)*. Edizioni di Storia e Letteratura.

Fink, J. L. (2011). Plethon's criticism of Aristotle's virtues: A note on *De differentiis*. *Greek, Roman, and Byzantine Studies, 51*, 483–497.

George of Trebizond. (1457). *Comparatio philosophorum Platonis et Aristotelis*.

Hladký, V. (2014). *The philosophy of Gemistos Plethon: Platonism in late Byzantium, between Hellenism and Orthodoxy*. Routledge.

Karamanolis, G. (2002). Pletho and Scholarios on Aristotle. In K. Ierodiakonou (Ed.), *Byzantine philosophy and its ancient sources* (pp. 253–282). Oxford University Press.

Masai, F. (1956). *Pléthon et le platonisme de Mistra*. Les Belles Lettres.

Monfasani, J. (1976). *George of Trebizond: A biography and a study of his rhetoric and logic*. Brill.

Plethon, G. G. (1973). *De differentiis* (B. Lagarde, Ed.). *Byzantion, 43*, 312–343.

Steiris, G. (2017). Pletho, Scholarios and the Arabic philosophy. In D. Searby (Ed.), *Never the twain shall meet: Latins and Greeks learning from each other in Byzantium* (pp. 339–358). De Gruyter.

Steiris, G. (2023). Trapezuntios and Bessarion on Arabic philosophy and science. In E. Anagnostou & K. Parry (Eds.), *The Neoplatonists and their heirs* (pp. 381–402). Brill.

Tatakis, B. (2003). *Byzantine philosophy* (N. Moutafakis, Trans.). Hackett.

Woodhouse, C. M. (1986). *George Gemistos Plethon: The last of the Hellenes*. Clarendon Press.

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 351 – 356

Book review

Nikolay Milkov. *The Tractatus, Wittgenstein's Method and Analytic Philosophy*. Palgrave Macmillan Cham, 2026, 312 p. ISBN: 978-3-032-09441-4

[Reviewed by ***Kamen Lozev****]

Was Wittgenstein an Analytic Philosopher?

One of the most interesting 'philosophical' stories which I was told some 35 years ago related to how the *Tractatus* had been translated into Bulgarian. Imagine in the middle of the 1980s an attic in Sofia where four people were regularly meeting for quite some time. Each of them was equipped with a copy of the *Tractatus* in German, English, French and Russian. N. Milkov[1] led the reading and the discussion of each and every single sentence of the *Tractatus* in the respective language, and then ended the discussion by offering the translation of the sentence into Bulgarian. The work had been tedious, but interesting and gratifying. This was the way, I was told, the small and 'enigmatic' book had been translated. After the *Tractatus*, Milkov also translated the *Philosophical Investigations* and the *Remarks on the Foundations of Mathematics*, and in 1988 the three main works of the earlier and the later Wittgenstein were published in a single volume in Bulgaria[2]. These translations paved Milkov's way into further investigations in the field of Analytic Philosophy, and if one looks at the list of his publications since the late 1980s one shall see that he is still committed to his initial interests in this kind of philosophy. All of this makes him particularly well suited to address issues concerning Wittgenstein from the perspective of an analytic philosopher. The title of Milkov's book – *The Tractatus, Wittgenstein's Method and Analytic Philosophy* – sounds slightly distracting since one expects that – as put in the middle – 'Wittgenstein's Method' will be the *link* attaching Wittgenstein to Analytic Philosophy. This expectation is not unsatisfied; however, after reading the book, the impression is that its gist has been more on the unity of Wittgenstein's method in his earlier and his later philosophy. As for the *link* to Analytic Philosophy, Milkov convinces us that Wittgenstein's work marks the pinnacle of Analytic Philosophy, narrowly understood as the one created by Frege, Russell and Moore. This is done at the end of the book, which makes me think that the last chapters are the most significant.

The book consists of three parts entitled, respectively, Influences on the *Tractatus* (3 chapters), Interpreting the *Tractatus* (5 chapters), and Wittgenstein's Method and

[1] **Nikolay Milkov** is Professor of Philosophy at Paderborn University, Germany.

[2] In 2003 there appeared another translation of Wittgenstein's *Tractatus* in a collection of translations, *Filosofia na logikata, ranna analitichna filosofia* (Philosophy of Logic, Earlier Analytic Philosophy), Sofia: St. Kliment Ohridski University Press, by Evgeni Latinov, but I think Milkov's translation remains to this day the standard one in Bulgaria.

*South-West University Neofit Rilski, Blagoevgrad; kamenlozev@abv.bg

©2026 Kamen Lozev. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

Analytic Philosophy (3 chapters). Milkov begins his book with a very clear statement: 'The task of this book is to present a valid interpretation of Wittgenstein's *Tractatus*' (p. 1). The book can, in fact, be seen as an embodiment of Milkov's conviction that, although the history of interpreting the *Tractatus* has generated a certain skepticism about whether it is possible to reconcile the "'different kinds of reading' of the Tractatus" (ibid.), the "'deep' analysis of Wittgenstein developed and applied in this book" (p. 2) enables us to "demonstrate that this skepticism is misplaced."

The title of my review replicates the title of Milkov's section § 12.1. In my view, *Was Wittgenstein an Analytic Philosopher?* stands as a *philosophical* question, at least for the fact that – like all philosophical questions – it is preceded by hard intellectual work: in this case it took Milkov 11 chapters before he could raise this question. Milkov needed to write these chapters, in fact the first two parts of the book, in order to substantiate his main conclusions in the final part of the book.

I have to emphasize here that in order to develop the content of his book Milkov uses his own 'method of "logical-contextual history of philosophy" introduced and developed by the present author' (p. 3), which in this case involves a close following of the texts and documents related to Wittgenstein and his works. Milkov also relies very much on 'the recent "archeological" studies of Wittgenstein's *Tractatus*, including those published online in (2016)' (ibid.). This shapes Milkov's writing as historical and systematic in character, and when reading the book, one feels as if it is a prolonged and true *demonstration* of this method, a deep investigation of Wittgenstein with exhaustive and meticulous references to all kinds of sources. It should be added, however, that Milkov's narrative is also developed as a kind of *constant dialogue* – in most cases a *critical* dialogue – with other Wittgenstein researchers, commentators and interpreters, which is only understandable in view of the task he sets himself to achieve, namely abolishing the fore-mentioned 'skepticism'. Severe criticism is especially levelled against the 'resolute readers', mainly in the face of Cora Diamond and James Connant but other famous authors like Brian McGuinness, Peter Hacker and Michael Dummett, to mention only a few, are not spared either.

Part I of the book is an overview of the strongest influences on the younger Wittgenstein from Russell, Frege and Tolstoy. It was due to them that the *Tractatus Logico-Philosophicus* was composed in the end, and they are widely discussed in any writing on Wittgenstein. The chapter impresses me with the careful study of the *reverse influence* which Wittgenstein himself exerted on his mentors, especially on Frege (the dramatic influence on Russell is well-known and is nicely presented here). Milkov, in section 3.8 investigates sources, which manage to convince us that 'Wittgenstein might well have exerted a significant influence on Frege' (p. 58), especially with his doctrine of 'showing', and especially in Frege's last years. Personally, for me, however, the most impressive sections in this first part of the book are the ones related to the profound investigations which Milkov undertakes with respect to Tolstoy's influence on Wittgenstein. (I somehow take it as if Milkov here is challenging me, showing *how little I have done* with my translation of the *Gospel in Brief*[3]

[3]Tolstoy, L. 2020. *Kratko izlojenie na Evangelieto* (Gospel in Brief), translated by K. Lozev, Blagoevgrad: Bon.

in my attempt to describe the 'religious Tolstoy'.) Tolstoy's influence, together with the influences of Weininger and Lotze, with their ideas of the unity of logic, ethics and esthetics, and the 'duty to oneself', Milkov uses to analyze the ethico-religious part of the *Tractatus*. Milkov widens the intellectual portraits of Tolstoy, revealing his theologo-philosophical views, his quest for truth and his critique of every form of oppression to convey the specific 'secularization' of the Gospel as Tolstoy interprets it: 'Ordinary Christians know what the message of Christ is – they understand it *immediately* and unconditionally. They just cannot express it. Tolstoy's objective on this score was simply to disclose this core truth' (p. 68, italics mine). In the discussion of Tolstoy's influence on the earlier Wittgenstein we learn that the more significant religious works by the Russian thinker were his translations from Greek of the four Gospels. We are used to thinking of Tolstoy's influence predominantly on the *earlier* Wittgenstein, but Milkov shows how the author of *War and Peace* had exerted great influence on the later philosophy of language of Wittgenstein in his *Philosophical Investigations*. We can agree with Milkov that it was while thinking about Tolstoy that Wittgenstein became aware, and in fact carried out, the so-called 'volition turn' in his second masterpiece. He writes in the *Big Typescript*, a virtual manifesto…:

> Tolstoy: the meaning (meaningfulness) of a subject lies in its being generally understandable. – That is true and false. What makes a subject difficult to understand –if it is significant, important – is not that it would take some special instruction about abstruse things to understand it. Rather it is the antithesis between understanding the subject and what most people *want* to see. Because of this the very things that are most obvious can become the most difficult to understand. What has to be overcome is not a difficulty of the intellect, but of the will.
>
> In short, Wittgenstein now realized that clarity is not enough. In philosophy, we also need a resignation of our emotions – the willful inclination to engage in philosophical thought and to use language triggered by a false understanding of the grammar of our language. To defeat this will, we need therapy – to be more exact, different therapies… (pp. 83–4).

Milkov puts a strong emphasis on this 'volition turn', discussing it in different contexts in the book, and it remains in the reader's mind as one of the most important messages.

The second and longest part of the book is a detailed recreation and interpretation of the *Tractatus*. The focus is put mainly on the ontological side of Wittgenstein's first masterpiece and its interpretation. It is here, in chapter 6, that we find the detailed and convincing criticism of the 'resolute reading' of the small book. In the course of Milkov's discussions in the chapter, one gets a deeper understanding of the main ontological concepts of the *Tractatus*, and especially how Russell and Wittgenstein used the concept of 'logical skeleton/scaffolding'. Here, Milkov exposes the ways in which resolute readers, and the New Wittgensteinians in particular, misinterpret the famous aphorism 6.54.: "My propositions are elucidatory in this way: he who understands me finally recognizes them as senseless, when he has climbed out through them, on them, over them. (He must, so to speak, throw away the ladder after he has climbed up on it.) He must transcend these propositions, and then he will see the

world aright." Milkov also shows how the Tractarian propositions should be viewed as a concept script in Frege's sense, taken to a higher level. These propositions help us to better see the world and the language we speak. This 'concept-script' is a tool which we are trained to use, and after understanding how to use it, it becomes useless and we can dispose of it.

Also, Milkov stands firmly against the attempts to view the *Tractatus* as 'therapeutic' in character. He convincingly writes: 'Among other things, while he may have read some Freud prior to finishing his work on the *Tractatus* in 1918, he initially adopted (not without reservations) the idea that philosophy has a therapeutic objective only in 1931. Contrary to the thesis fielded by Conant and Diamond, we hold that the *Tractatus* advances no program for therapy but rather argues, as we've seen, for an activity designed to teach its readers better to understand their language, and by that means trains them to think with greater cogency. In this way Wittgenstein subscribed to the classic Cartesian view that the vocation of philosophy is to cultivate clear and distinct thought' (pp. 139–40).

In this second part of the book, Milkov, sticking mainly to the *Notebooks of 1914–16*, the *Prototractatus*, and other extant evidences, is able to recreate the origination of the central ideas of the *Tractatus*, locate the places, and sometimes the dates, together with the exact circumstances and the factors which led to their difficult formulation in Wittgenstein's mind. This helps the reader to deeply penetrate into the ways by which the young soldier, and later officer of the Austrian Imperial Army, struggled to create his ideas, then chains of ideas, and finally the 'tree' of his first book.

The third part of Milkov's book, in my view, is the most important one, since it is precisely here that Milkov displays his understanding of Wittgenstein's method and how this method creates the unity of Wittgenstein's two philosophies, as well as their relation to Analytic Philosophy. In fact this 'display' is the readers' 'expectation' encoded in the very title of the book. However, in the beginning of the book, Milkov hints at what "Wittgenstein's method" is with the brief pronouncement that '…the objective of the present book is to reveal Wittgenstein's deep insight that has a form of a method' (p. 2), but here, in Part III, the method is fully justified and examined in detail, presented "in action" – both as practiced by Wittgenstein himself and his contemporaries and within the history of philosophy.

The first of the three chapters of Part III, Chapter 10, is a kind of preparation for the analysis of Wittgenstein's method. Milkov urges us to recollect that in 1943 Wittgenstein pleaded to the Syndics of Cambridge University Press to publish in a single-volume edition the *Tractatus* and the *Philosophical Investigations* (he even coined the common title *Philosophische Untersuchungen der Logisch-Philosophischen Abhandlung entgegengestellt*), believing that 'making explicit the many points the two works have in common…could aid readers in properly understanding his thinking in the *Philosophical Investigations*' (pp. 218-19). The project never materialized; however, Milkov tells us what the results would have been in case it did. In this chapter he enumerates and discusses the dissimilarities (six in total) and similarities (fourteen in total) between the *Tractatus* and the *Philosophical Investigations*. He then reviews

the theses of *Philosophical Investigations* which can be shown to be further developments of ideas formulated in the *Tractatus*. Through this short comparative review Milkov convinces the reader that Wittgenstein's *modus operandi* is in fact 'the singular *skill* with which he carried out his "philosophical work"' and that it had 'remained practically unchanged through all periods of his philosophical development' (p. 235). This *skill*, obviously, creates the unity of everything he ever wrote.

In Chapter 11, Milkov further investigates the issue of Wittgenstein's method. He observes that while 'in the *Tractatus* Wittgenstein speaks freely about "my method" (4.1121) and about the "correct method in philosophy" (6.53)', in his 'Cambridge lectures of the early 1930s, as recorded by G. E. Moore, Wittgenstein repeatedly referred to his "new method", one that makes it possible "to reduce philosophy to a matter of *skill* ...[in] which we have to be *trained*"' (p. 239). Now it seems as if Wittgenstein's method is conclusively identified, but Milkov still remains unsatisfied and further asks the question: 'But what exactly did Wittgenstein mean by "method"?' (p. 240). At this point he turns to the analyses of Ioachim Schulte on this theme – borrowing two of Schulte's ideas and also criticizing him for taking 'the same misstep as do the resolute readers of Wittgenstein' (p. 241). The first idea is that 'instead of directly pressing the issue of Wittgenstein's method, we do better by concentrating on his *skill* in examining philosophical problems' (p. 240). Schulte is aware that, unlike the widely-spread notion that "method is theory in action", 'Wittgenstein had no explicit theory' (ibid.), and therefore cannot be approached in a standard way. The second idea – not unknown to us already – is 'Schulte's claim that we are better placed to understand what Wittgenstein called his method if we view his philosophical work as a specific *skill* that he put in use, or recommended that his readers exercise, when treating philosophical problems' (p. 241). Schulte's analyses, however, relate to the 'later Wittgenstein', which does not *fully* serve Milkov's thesis that Wittgenstein's method remains the same in both his philosophies. Milkov, therefore, devotes section 11.4 in discussing Wittgenstein's philosophical skill first in his early and then in his later philosophy.

Wittgenstein's extraordinary philosophical talent and profound insights help explain both why he "wrote his philosophical remarks in sketchbooks" and why his philosophical reasoning took such an aphoristic form. Attempting to explain the uniqueness of this *modus operandi*, Milkov turns to Wittgenstein himself and his statements: 'Nearly all my writings are private conversations with myself', 'working in philosophy ...is really more a working on oneself; on one's own interpretation; on one's way of seeing things' (p. 252). Milkov, however, strongly emphasizes that applying one and the same method in philosophy does not mean arriving at the same ideas. Wittgenstein's case was exactly this: the application of his philosophical skill in the *Tractatus* drove him to the ambition of achieving the 'logico-philosophical "truth" once and for all and in completely definitive terms', while in the *Philosophical Investigations*, he came to the conclusion that philosophical propositions are 'perennially open-ended' and ones 'which we perform *ad libitum*' (p. 246).

In the last chapter of his book Milkov traces the evolution of what had started as 'early Analytic Philosophy' with Frege, Russell and Moore into the so-called 'later

Analytic Philosophy'. I think Milkov here dispels the myth of a certain 'analytic-continental split' in contemporary philosophy demonstrating that the 'analytic way' or 'analytic trend' in philosophy originated as a natural development of Germanophone and Anglophone philosophy. The book ends with a discussion, or rather 'refinement', of Peter Hacker's general position 'that Wittgenstein was the pinnacle of analytic philosophy' (p. 278).

In conclusion, Professor Milkov's book should be read as a historical-systematic work. It is undoubtedly an enlightening "journey" through Ludwig Wittgenstein's oeuvre, with major stops at his most important works, which Milkov meticulously "dissects" in order to identify the method employed by their author. In this sense, the book truly "enlightens" the reader. However, to understand it fully, one must be well acquainted with the thinkers and works cited and criticized throughout the book. Milkov certainly assists the reader with quotations and detailed analyses, yet the book nevertheless presupposes substantial prior knowledge. In this respect, it is not a book for beginners in Wittgenstein studies or Analytic philosophy.

Although readers like me may find little of substance to disagree with in the book, I expect that it will nevertheless attract opponents and serious critics, given the many significant issues and aspects of Wittgenstein's work that it addresses. Yet, as Popper reminds us, thoughtful and well-intentioned criticism is perhaps the best response that a work of genuine merit, such as the one reviewed here, could hope for.

Balkan Journal of Philosophy (2026), vol. 18, issue 2, 357 – 361

Book review

Stavros Ioannidis and Stathis Psillos. *Mechanisms in Science: Method or Metaphysics?* Cambridge: Cambridge University Press, 2022, 250 p. ISBN: 978-1-316-51990-5

[Reviewed by ***Aida Roige**** and ***Saúl Pérez-González*****][1]

What do scientists mean when they invoke mechanisms? Are mechanisms features of the world to which scientists are ontologically committed, or are they primarily tools for organizing inquiry and explanation? Stavros Ioannidis and Stathis Psillos take these questions as their point of departure in their book *Mechanisms in Science: Method or Metaphysics?* (2022), which provides an original perspective on mechanisms in science.

Drawing on a careful examination of how scientists understand and use the notion of mechanism in practice, Ioannidis and Psillos argue that the notion's primary role is methodological rather than metaphysical. Their central claim is that the concept of mechanism is a methodological device within scientific investigation, and thus our account of mechanisms must accordingly be ontologically non-committal. They defend the metaphysically deflationist view they call Causal Mechanism (CM): "(CM): A mechanism is a causal pathway that is described in a theoretical language" (p. 3)

The scientific commitment to mechanisms is, thus, a commitment to describe certain causal pathways (or processes) using mechanistic methods and terms – and not a commitment to the world being a certain (metaphysically thick) way. In this respect, Ioannidis and Psillos depart from what has become the common ground among many mechanistic accounts, which tend to adopt a more realist stance. The more persuasive their arguments, the more they appear to shift the burden of proof onto mechanists who are not agnostics in their ontological commitments, making this an original and thought-provoking book.

In this review, we aim to present, analyze, and reflect on the book's main contributions. We proceed as follows: in the next section, we provide a part-by-part overview of the book, and in the final section, we critically assess its principal arguments and implications.

1. Overview of the book

The book is divided into three parts: *Ideas of mechanism*, *Causation and mechanism*, and *Beyond New Mechanism*. The first part, which includes chapters 1 and

[1] Equal contribution; authors' names are listed in random order.

*Department of Humanities: Philosophy, Language and Literature Theory, Universidad Carlos III de Madrid; aroige@hum.uc3m.es ** Department of Philosophy, University of Valencia; saul.perez@uv.es]
©2026 Aida Roige and Saúl Pérez-González. This article is distributed under a Creative Commons CC BY-NC 4.0 license.

2, addresses the historical development of the concept of mechanism. Ioannidis and Psillos discuss the relationship between the Old Mechanism, advocated by authors such as Descartes and Boyle, and the New Mechanism. They argue that, although both approaches involve metaphysical theses and a methodological theory, they diverge regarding which is considered most fundamental. The New Mechanism, unlike the Old Mechanism, starts by studying mechanisms as a concept-in-use in science and aims to base the ontology of mechanisms on it.

The second part, which includes chapters 3 to 7, develops Ioannidis and Psillos's Causal Mechanism. The authors hold that CM involves three main theses: (1) mechanisms are causal pathways, (2) causal relations among components are to be viewed in terms of difference-making, and (3) metaphysical agnosticism. CM, unlike dominant accounts in the New Mechanism, does not involve additional ontological commitments. They argue that mechanisms should be described in the theoretical knowledge of the field, not in some privileged language of ontological categories. To develop and argue in favor of CM, the authors address in detail the identification and study of the biological mechanism of apoptosis. They examine the cytological description of the mechanism of apoptosis to identify the main features of mechanisms in biology.

In this part of the book, Ioannidis and Psillos also address the relationship between mechanisms, causation, laws, and counterfactuals. They argue that laws and counterfactuals are crucial to understand within-mechanism interactions. The authors also introduce the distinction between "mechanism-for" and "mechanism-of". A mechanism-for is a mechanism that is responsible for a certain phenomenon or underlies a certain function, while a mechanism-of underlies or constitutes a causal process.

The remainder of the second part of the book goes on to defend (2), the thesis that causation (and hence causal mechanisms) is best understood as underpinned by difference-making relations (a kind of counterfactual dependence). This is counterposed to the (Aristotelian) tradition of understanding causation in terms of production and causal powers. Ioannidis and Psillos argue for (2) on several different fronts: for example, they point out that it is hard to make sense of productive claims without assuming the truth of counterfactual claims, which would seem to suggest that contrary-to-fact conditionals are metaphysically more fundamental; furthermore, they point out that the Woodwardian interventionist approach to causation (upon which the mechanistic one relies) is based on counterfactuals. Regardless of whether one agrees with their take on these issues or not, they do a wonderful job of examining mechanistic activities critically and situating them within the broader framework of discussions about causation.

The last part of the book is dedicated to defending (3): metaphysical agnosticism; that is, refusing to commit to a particular take on causation or to the reality of constitutive relations (something that most accounts of mechanisms do). To do so, Ioannidis and Psillos first question the existence of constitutive mechanisms as described by Craver (2007), namely those underlying S's Ψ-*ing*. This is because mechanisms sometimes have external components, and because some mechanisms may exist outside the entity whose behavior they underlie. Thus, if the components do not materially

constitute the phenomena, it is not clear in what sense they are part of the constitutive mechanism for something.

This has repercussions for how we understand Craver's notion of "levels of mechanisms." Here, Ioannidis and Psillos recall (and defend) the traditional distinction between levels of composition and levels of explanation. They defend the view that whatever contributes to the phenomenon is part of the same causal pathway, and that causal pathways may involve entities at different levels of composition. Thus, multilevel explanations are causal explanations involving entities at different levels of composition.

Lastly, Ioannidis and Psillos go beyond CM – the claim that mechanisms are causal pathways described in theoretical language and underpinned by networks of difference-making relations – and embrace a thesis called Methodological Mechanism (MM). According to MM, the notion of mechanism (as a concept-in-use in scientific practice) is employed methodologically, that is, for the purposes of investigation and explanation. They take the scientific commitment to mechanisms to be ultimately a methodological commitment to seek a certain kind of explanation, and argue that inflationary accounts of mechanisms should not regulate scientific practice.

2. Critical assessment

One of the main positive contributions of the book is a detailed analysis of the historical predecessors of contemporary mechanical philosophy; this will definitely delight the historically-inclined reader and places the new mechanistic philosophy in the broader context of traditional philosophy of science. This is a notable contribution, for while it was commonly assumed that there is a connection between the new mechanistic philosophy and the Old Mechanism, this relationship has often been presented in a superficial way. In contrast to this tendency, Ioannidis and Psillos provide a detailed reconstruction of classical mechanistic proposals, as well as the main debates surrounding them. The authors consider that this is important not only for illuminating the relationship between the Old and the New Mechanism, but also for the assessment and development of the latter. For example, based on that analysis, we understand better the distinction between mechanism as a metaphysical concept and as a concept used in science.

However, it is not entirely clear to what extent the analysis of the Old Mechanism can provide us with adequate guidance for understanding and developing the new mechanistic philosophy. This is because of several reasons: first, it is not clear to what point the debates surrounding the Old Mechanism have informed the development of the New Mechanism. At times, mechanists seem to have even purposefully distanced themselves from mechanical views from the past; for instance, when discussing the terminological choices of Machamer, Darden, and Craver (MDC), Peter Machamer says that they chose the term "entity" because it "seemed to carry fewer historical and philosophical presuppositions than 'substance'" (Machamer, 2004, p. 27). Second, although some ontological commitments adopted by the Old Mechanism proved problematic, ontological neutrality (the stance towards which Ioannidis and Psillos

lean) may not be adequate either. Ontological considerations about mechanisms can, for instance, play a relevant role in guiding the identification and characterization of mechanisms in scientific practice (Craver & Darden, 2013), proving thus useful.

The main proposal of the book is Causal Mechanism. According to this account, a mechanism is a theoretically described causal pathway, which in turn is underpinned by networks of difference-making relations. This proposal, as it is usual in the New Mechanism, is developed taking scientific practice as reference. Authors are interested in mechanisms as a concept-in-use in science. Moreover, in line with recent trends (see, for instance, Glennan 2017), they aim to provide a general and encompassing characterization of mechanisms. They aim to account for mechanisms across the sciences. Nonetheless, CM also seems to face limitations. It is unclear whether CM, which is developed taking some paradigmatic biological examples as reference, can be generalized and extrapolated to other scientific fields. Given the diversity of mechanisms across the sciences, this kind of strategy usually faces difficulties for developing general definitions of mechanisms (P´erez-González, 2019). Moreover, given its minimalism and ontological agnosticism, its value for scientific practice and philosophical reflection may be limited. In this sense, for example, it seems ill-equipped to account for the use of mechanisms in areas such as sociology, where ontological commitments about the components of the social mechanism play a relevant role (P´erez-González, 2020). A similar concern arises for cognitive science, where ontological assumptions go hand-in-hand with the discovery of different *sorts* of mechanisms in the mind/brain (Roige, 2025).

One of the book's central themes is, as noted above, the metaphysics of mechanisms. The book develops a careful examination of the ontological dimension of some of the main characterizations of mechanisms, situating mechanistic views within broader debates in the philosophy of science. In doing so, the authors make the book an especially valuable resource for those seeking to locate discussions of the new mechanistic philosophy within the wider theoretical landscape.

The new mechanistic philosophy, however, is not a single view, but a family of theories that have important differences among them. Ioannidis and Psillos take on the non-trivial task of trying to find the common ground between these views and present them as somewhat of a unified account. Nonetheless, this argumentative strategy carries important risks with it: the resulting rendition may be a somewhat simplified picture of the new mechanistic literature, conflating positions advanced by different authors that were not intended to be unified in that way. In this sense, it is unsurprising that they identify tensions and inconsistencies (e.g., in claims made about mechanisms' functions, without specifying whether they are talking about proper or causal role function), since the claims in question often belong to distinct theoretical frameworks. It may be argued that some of the difficulties they highlight arise only once heterogeneous views are treated as a single doctrine.

In summary, *Mechanisms in Science* is a valuable and rigorous contribution that fills a gap in the debate. The authors integrate the historical analysis of mechanistic

philosophy with contemporary debates on the metaphysics of mechanisms. Moreover, they introduce a minimalist, ontologically agnostic characterization of mechanisms that contrasts with popular views in the new mechanistic philosophy, putting pressure on them as to why they should make more ontological commitments than strictly necessary to account for scientific practice. This book is of interest to anyone interested in the new mechanistic philosophy, especially those interested in its metaphysical dimension.

Acknowledgments: Work leading up to this review was funded by the project "New approaches to evidence: theoretical and practical developments", financed by MICIU/AEI/10.13039/501100011033 and by ERDF/EU (PID2024-155508NA-I00), and, in Roige's case, by the project "Cognición Mecanicista: ¿es posible integrar las ciencias de la mente y la computación?", financed by the Programa Propio de Investigación de la Universidad Carlos III de Madrid (Ayudas para la Actividad Investigadora de los Jóvenes Doctores, 2025/00796/001).

References

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Clarendon Press.

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. University of Chicago Press.

Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.

Ioannidis, S., & Psillos, S. (2022). *Mechanisms in Science: Method Or Metaphysics?*. Cambridge University Press.

Machamer, P. (2004). Activities and causation: The metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science*, 18(1): 27–39.

Pérez-González, S. (2019). The Search for Generality in the Notion of Mechanism. *Teorema*, 38 (3), 77–94.

Pérez-González, S. (2020). Mechanistic explanations and components of social mechanisms. *European Journal for Philosophy of Science*, 10, 35.

Roige, A. (2025) How is the Mind/Brain Composed? Two Sorts of Mechanisms in Cognitive Science. *Review of Philosophy and Psychology*, 16, 1427–1450. https://doi.org/10.1007/s13164-025-00781-0