BALKAN JOURNAL OF PHILOSOPHY

Vol. 17, Issue 2, 2025

The issue is sponsored by the Bulgarian National Science Fund, competition "Bulgarian Scientific Periodicals – 2025", project No: КП-06-НП6/61 from 05.12.2024.

TABLE OF CONTENTS

Balkan Journal of Philosophy (2025), vol. 17, issue 2, 110–123

THE GRADATIONAL SENSE OF MECHANISMS

Predrag Šustar*

Abstract: The famous Dobzhansky's (1973) dictum, according to which "nothing in biology makes sense except in the light of evolution", is supposed to involve all the biological sciences. But, what about biological mechanisms, such as, most notably in the philosophical literature, the mechanism of protein synthesis? If the dictum applies here as well, how should we correctly account for its evolutionary-historical character? Recently, Garson (2023) has argued that whether or not something is a biological mechanism partly depends on its history. In this paper, I examine how the mechanism of protein synthesis can evolve from accidental and unstructured causal processes occurring in the genome (see Van Oss and Carvunis 2019). Such historical analysis allows me to outline some general features of an important type of biological mechanism. It shows that such features can be discerned without invoking the mechanism's evolutionary history in the way argued by Garson's current account, and that the sense of mechanism in question is not typical for biology.

Keywords: thick mechanisms, thin mechanisms, evolutionary history, etiological functions, artifact functions, machine analogy, human genome, *de novo* genes and proteins.

1. Introduction

Recently, Garson (2023) has argued that biological mechanisms are historical entities. By this, he does not mean just that they are results of historical processes, but that whether or not something is a mechanism, and what it is for, partly depends on its history. In this paper, I explore what insights about biological mechanisms can be drawn by looking at their history, i.e., how they are emerging, according to some recent biological theories and their explanatory models. Specifically, I look at the history of one of the canonical biological mechanisms – protein synthesis, and examine how the emergence of new functional genes from noncoding genomic sequences is explained.

My case study is concerned with the theory of *de novo* gene synthesis (see Tautz and Domazet-Lošo 2011; Van Oss and Carvunis 2019; Weisman 2022; Peng and Zhao 2024). It needs to be emphasized, however, that the mechanisms in question do not arise from scratch. Rather, their emergence relies on the existing cellular machinery responsible for the gene expression of the already present, "old" genes. I focus on the transition, enabled by natural selection, from accidental causal processes into a more structured process that can be considered as a mechanism. A key feature, I claim, ensuring this transition, is buffering from other causal processes going on in the genome.

Tracking the transition under consideration, in turn, allows me to draw some general features of an important type of biological mechanism, exemplified here by the mechanism of protein synthesis. My analysis yields significantly different results in comparison to Garson's

*Department of Philosophy, University of Rijeka; psustar@uniri.hr

(2023), even though I agree with him that looking at mechanisms' history is informative in interesting ways. But, I show primarily that whether or not something is a prominently biological mechanism basically does not depend on the corresponding evolutionary history. Here I give precedence to the standard New Mechanist ahistorical approach in identifying a biological mechanism, and what it is for, according to which, roughly, all we need to do is: observe it, take it apart, manipulate its constituent component-parts, and see how the outcome is produced (see, in particular, Craver and Darden 2013). However, while New Mechanists have distanced themselves from the idea that mechanisms are machines (see Craver and Tabery 2023), I take the machine analogy seriously, especially for a certain type of biological mechanism, as shown in Section 4. The analogy, nevertheless, is not with machines intended as *systems*, that is, in the sense of material entities, such as a mousetrap. Rather, the analogy relates to causal *processes* themselves, occurring, however, in some material bearer. For example, the flow of genetic information is a causal process, but it is not the same when occurring in different material bearers: DNA or RNA macromolecules.[1] I think that determined biological causal processes have features similar to causal processes going on in the paradigmatic machines, such as, most notably, clocks, as I argue in Sections 4 and 5.

This means that my notion for this kind of biological mechanism, unlike Garson's (2023), is not characteristic or typical for biology, even if it is closely tied to the notion of function. The notion of function in question is a relatively broad one, capturing the idea of what a mechanism is designed to do (see Kitcher 1993). In the case of biological mechanisms such as protein synthesis, natural selection takes the role of a designer, while in the artifact case, we need to invoke some kind of elaborate human intention. One potential benefit of such an approach is that it does not need a separate notion of mechanism for the overall area of synthetic biology and biological engineering.

Finally, my analysis also shows the importance of a multitude of accidental causal processes occurring in the genome, some of them resulting in interesting mechanistic arrangements, as illustrated by the *de novo* gene emergence explanatory models in Section 3. Such mechanisms in a *thin* sense (see Garson 2023) can, at least derivatively, be considered as historical entities in the sense proposed by Garson's account.[2] Depending on the context

[1] In that regard, my view is most sympathetic to Krickel's (2018) corresponding account, according to which mechanisms are *entity-involving occurrents*, thus, advocating a moderate processual interpretation of the abovementioned issue. In other words, on my reading, Krickel's account does not ignore the system or material bearer feature and its influence on the respective causal process, despite holding firmly onto a basically processualist view. In that sense, I am more moderate than Garson's fully processual interpretation of biological mechanisms.

[2] In that regard, it is worth noting that Garson argues for the so-called "weaker claim" about the historical character of prominently biological, thick, mechanisms – in other words, the claim according to which whether something is a mechanism in the above sense, and what it is for, depends on having the right history. The claim openly does not specify what kind of history is needed here. (Ultimately, that depends on the function notion we choose.) Thus, thick mechanisms are not necessarily adaptations; that would be a 'stronger claim', I suppose. My account, which will argue for a continuum between different types of biological mechanisms, will however claim that, at least, thick mechanisms are adaptations, given the role played by natural selection in the progression of events leading to a new functional gene, as shown in Sections 3 and 5, more specifically.

they occur in, and a specific sequence of events preceding and following them, I will classify a certain relatively unstructured causal process as a thin or 'in between thin and thick mechanism' only if it leads to more structured mechanistic arrangements.

In the next section, I clarify in more detail the abovementioned senses of the mechanism notion that I discuss throughout the paper. In Section 3, my case study of how new functional genes can evolve from noncoding genomic sequences is described, and the main features of one type of biological mechanism are outlined. Section 4 shows that the machine analogy is appropriate for a type of biological mechanism, such as protein synthesis. In conclusion, in Section 5, I discuss what general lessons about biological mechanisms can be drawn by examining their evolution.

2. Some preliminary distinctions

I will follow Garson (2023) in distinguishing two main senses of the mechanism notion, but allowing that there may be other senses, as well.[3] In the first sense, which he calls "thick" or "mechanisms in the prominent biological sense" (Garson 2023, 921), in order for something to be a mechanism, it needs to have a proper function.[4] Such mechanisms, he claims, are composed of different parts, which, when aligned properly, bring about the corresponding phenomenon with some degree of regularity. But, most importantly, they are also the kinds of things that can break (see also note 4, feature (3)). Such mechanisms, according to Garson, have been inappropriately associated with machines. The difficulty with this association is, he claims, that it is unable to capture the abovementioned process-like, dynamic nature of biological mechanisms.

In what follows, I will refer to this as the *thick* sense of biological mechanisms, since Garson's account states that there is another "thin" sense. This sense of a mechanism basically refers to any causal process or trajectory.

My main objective in this paper is to deal with mechanisms in the thick sense, and offer their characterization by examining how they evolve in molecular biology from the thin sense; more specifically, from accidental causal processes. I start out by taking protein synthesis, generally used as a faithful instance of a biological mechanism (see, in particular, Machamer, Darden and Craver 2000). I am not focusing specifically on its ordinary, everyday functioning, but, rather, on the transition from a non-structured causal process to a token-mechanism, as I will clarify shortly. The idea is to pinpoint important changes that take place and allow us to identify a newly arisen causal structure as a (thick) biological mechanism.

[3] At the end of the paper, I hint at the possibility of another sense that might be relevant to the biological sciences, but a more elaborate analysis of this point goes beyond the scope of the present paper.

[4] Proper functions are usually characterized by the following three important features: (1) a trait's function should be useful to the individual organism bearing the trait; (2) by attributing a proper function to the trait, we are thereby explaining why the trait is present in a population; and (3) the function bearing traits can malfunction or be dysfunctional. Note that (2) is tied to the etiological account of biological functions. Garson, however, in the corresponding account refrains from a commitment to a specific account of functions, and adds that *at least sometimes* attribution of a function to a trait is also an explanation of its existence (see Garson 2023, 925). I will not enter here into specifics of this debate, but will briefly address the relationship between functions and mechanisms in Section 5.

In the final section of the paper, I additionally touch upon the question of how to characterize mechanisms in the thin sense. I will argue that there is room for delineating a certain kind of relatively loose and unstructured mechanism as a properly biological mechanism in this sense, contrary to Garson's dichotomous classification. Namely, I offer reasons for thinking that such mechanisms cannot be equated with just any causal trajectory.

3. Protein Synthesis under Construction

Protein synthesis is one of the most fundamental biological mechanisms by which individual cells build themselves up – roughly, by which, the containing organism builds its endophenotype and phenotypic characteristics, more generally. As such, it has been shaped by natural selection into a complicated 'machinery', especially in the eukaryotic domain, ensuring that it proceeds through a specific sequence of events. Finally, its outputs are obtained in a regular manner.

I will try to answer the question of how a new token-mechanism of protein synthesis can evolve, and ask at which point it is safe to assume that a full-blown new mechanism has been co-opted by the extant biomolecular and cellular mechanisms. To get some insight into this issue, I examine the ways in which new genes originate, as pointed out above, in already existing genomes. In that regard, two main answers have been offered to one of the most important questions in molecular and evolutionary biology, namely, 'how genes emerge and become functional' (see Tautz 2014). A chronologically first approach to that question explains the phenomenon of a gene emergence via *duplication and divergence* processes. A more general approach views genes as arising *de novo* from sequences that were not previously coding segments, on which, thus, I focus in the present paper. The duplication and divergence approach and corresponding explanatory models have been, until fairly recently, the only candidate explanation for the above question. The approach claims, in brief, that a putative new gene emerges from the gene ancestor by the duplication process (see, e.g., Schmitz and Bornberg-Bauer 2017). As the new gene duplicate is operating, so to say, 'in a shadow' of its gene ancestor with respect to selection pressures in the genome, it can more readily diverge and end up: (1) as a pseudogene; (2) it can sub-functionalize a part of its ancestor's original function; or (3) it can neo-functionalize, that is, acquire a new function within the containing gene family. Put differently, gene duplicates are free to diverge and undergo quite radical changes in their genomic sequences. As a consequence, the similarity to the founder gene is lost and, thus, it is classified as an "orphan" gene.

Accordingly, I focus on a more general scientific explanation of the origin of "orphan" genes, i.e., genes that do not have detectable homologues in other lineages, most frequently, in other species (see, e.g., McLysaght and Guerzoni 2015). Such "orphan" gene sequences can constitute up to one third of genes in all eukaryotic genomes (see Tautz and Domazet-Lošo 2011). In this paper, I thus foreground the biological theory and its explanatory models concerned with *de novo* gene origins, which, roughly put, are about the ways in which an initially non-coding genomic sequence can develop into a fully functional new gene (see Carvunis et al. 2012; more recently, Weisman 2022).

Let me first describe more general steps in the evolution of new token-mechanisms of protein synthesis, which originate from intergenic, noncoding DNA segments. Subsequently, I will give a more detailed insight into the overall process here under consideration.

As already emphasized, the whole process starts from a genomic sequence classified as 'intergenic and non-coding. Now, a *first* step leading to a *de novo* gene, according to a leading explanatory model (see Figure 1 below), refers to a non-coding genomic segment that is transcribed at low levels. A *second* general step of a *de novo* gene emergence refers to the process of translation. Both abovementioned occurrences are not unusual in the functioning of a genome, as the explanatory models pertaining to this biological theory frequently point out when addressing the way in which eukaryotic mechanisms of protein synthesis work.[5] Many non-protein coding RNAs can be found in ribosomes, from which we can infer that a certain number of superfluous transcripts get, nonetheless, translated into rudimentary proteins (see Figure 1, the three columns from the left). Finally, in a *third* general step of the overall process, transcripts become translated at higher levels, and the resulting peptides gain determined structural features and, in a scenario favorable to a containing biological system (ultimately, an individual organism), start to provide adaptive advantages. At this point, the fortuitous outcomes of the mechanisms of transcription and translation, involving antecedently non-coding genomic segments, become much more stable occurrences that can be fixed in a population via natural selection. In that case, as shown in Figure 1, a new functional gene has been produced.[6]

Now, before pausing on the evolutionarily most significant steps within the process, and examining how these and eventually other steps of *de novo* succession relate to Garson's "core argument" for the historical character of prominently biological mechanisms, let's take first a more fine-grained look at the overall process.

According to the *de novo* theory, an intergenic genomic sequence can eventually develop into a "proto-gene" via a short preexisting "Open Reading Frame" (ORF), as shown by Figure 1. In short, proto-genes are those genomic entities that exhibit stable expression and/or translation, but do not amount to having a *proper* function, or in other words, a "minimal" function, that is, a function as a mere effect of a certain trait (For the way in which this traditional distinction from the functions debate is used in approaching the mechanism-function relationships, see most recently Garson 2023.) Let us now examine what the authors of the models of new genes emergence have in mind when they state that some level of proto-functionality can emerge at the genomic level without amounting to a proper function, and how a transition to the proper functionality may nevertheless occur.

The progression in question proceeds in a stepwise manner. As illustrated in Figure 1, the process begins with a "junk" genomic segment transcribed at low levels. Such transcription is a normally-occurring process in large portions of the genome, even non-coding ones. In the next step, this segment can acquire a short "open reading frame" (ORF), i.e.,

[5]Large amounts of noncoding DNA packed in the genomes are subject to low levels of "pervasive" and "promiscuous" transcription and, to a lesser degree, translation, even though these genomic segments exhibit no apparent bioactivity, and are not under detectable selective constraints (see Clark 2011; Ingolia et al. 2014).

[6]There are two leading explanatory models based on the biological theory in question, which further examine some of the crucial difficulties in the abovementioned general succession of events: the "RNA first" and "ORF first" explanatory models, as pointed out in Figure 1. Although in what follows I will describe more specific steps in the *de novo* overall process, I will nevertheless set aside the differences between the two explanatory models, as well as their respective shortcomings. In my view, these issues go beyond the scope of the present paper. (For more on this thread in the scientific literature, see Weisman 2022.)

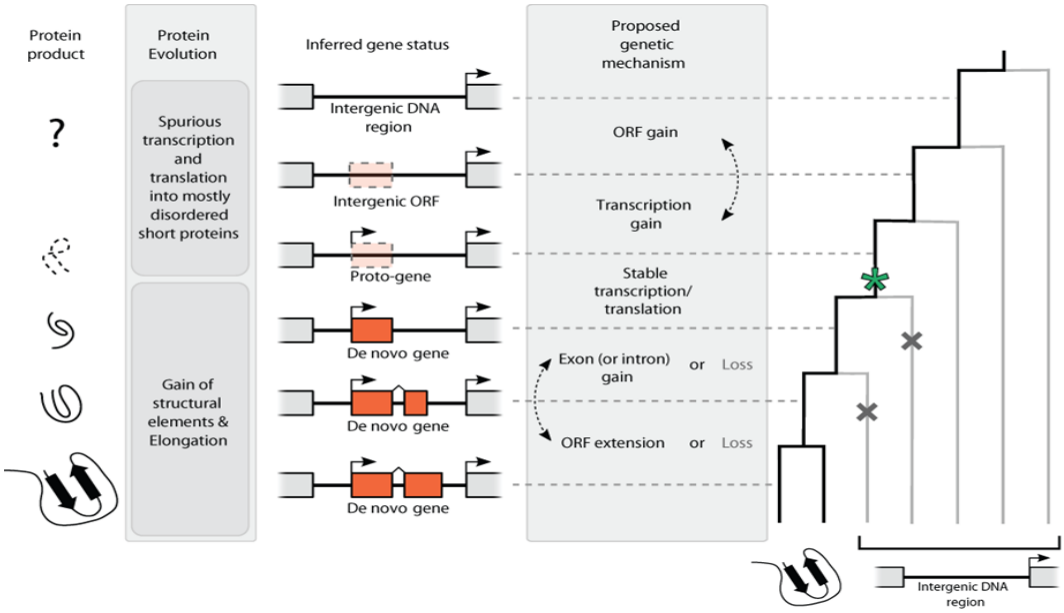

FIGURE 1. Schematic depiction of *de novo* protein-coding gene emergence (from Schmitz and Bornberg-Bauer 2017).

a continuous stretch of codons with a start codon and a stop codon, that can be translated into a peptide. The translation occurs at the next step of the progression, and a short ORF can thereby result in some rudimentary proteins. This step is still a part of the fortuitous goings on in the genomic environment – more specifically, in the ways the mechanism of protein synthesis is operating within a cell.

Even though the above step is usually classified as a proto-gene, I think that the proto-functionality may occur only at a subsequent step. I agree with the general approach of the authors of the shown *de novo* gene birth model that there is an evolutionary continuum ranging from non-genic sequences to genes themselves. Given such a continuum, perhaps it is unnecessary to exactly pinpoint a stage in which proto-functionality emerges. In this step, there is no ground, even minimal, for distinguishing accidents from emerging functionality. A further step, however, displays some interesting new features that can be associated with proto-functionality.

The ORF in question increases in length, which leads to an increase in the level of expression. The increase in length can happen through a loss of stop codons, by a shift of start codons or by a duplication followed by a fusion with other ORFs (see Carvunis et al. 2012, 372). What is interesting for the case of emerging proto-functionality is that the increase in the level of expression of the ORF may happen via the recruitment of existing regulatory elements. This incorporation into the existing regulatory networks starts to ensure stable expression. The new proto-gene becomes transcriptionally active, i.e., it acquires a promoter and gets co-regulated with genes it interacts with or those that are expressed in the same life stages or environmental conditions. The corresponding proto-gene's transcript is translated at higher levels, and the resulting peptide starts to provide an adaptive advantage to, ultimately, the containing individual organism. At this point, the structural profiling

of proto-genes can begin: the secondary structure of proteins exhibits more stability, even though this process is slower than the gain of abovementioned regulatory elements. Thus, this step might be crucial in establishing proto-functionality of genomic segments, because it exhibits features ensuring the stability of expression. The stability is secured by exploiting extant regulatory networks that are under selective pressures. However, the stability of a (proto) gene expression is insufficient by itself for proto-functionality, which also requires a positive contribution to the fitness of the containing individual organism, that is, an adaptive advantage. A final step in the overall *de novo* progression of events specifies that a new functional protein product is preserved by natural selection, securing in this manner that the adaptive end product of a newly coopted token-mechanism of protein synthesis is regularly and stably produced. Additionally, the corresponding new genomic segment gets spread across different taxonomic groups, thus ceasing to be an "orphan" gene. (For yet further, more specific, steps in this progression, see most recently Peng and Zhao 2024.) Let's see now how both general and the previously described specific steps of the *de novo* explanatory models connect to Garson's attempt to explicate a historical character of prominently biological mechanisms, such as, paradigmatically in the New Mechanist tradition, the above mechanism of protein synthesis.

Thus, I investigate how a relatively loose causal process acquires structure and organization that will ensure that it produces a certain outcome with a high degree of regularity (consider again the initial step in Figure 1 and related description). An important feature of the whole process is a relatively rapid integration of the new gene into existing regulatory networks, acquisition of regulatory signals controlling gene expression, and stable localization into a specific subcellular compartment. (See, for instance, Abrusan 2013.) Integration within the existing regulatory networks is important, because it ensures that all the steps in the gene expression process can be modulated, thereby allowing the control over when, where, and how much of the gene product will be produced in the cell. This, in turn, gives a clear structure and organization to the casual process leading up to the functional protein outcome. In other words, the process of production of the corresponding protein is no longer an accidental matter. Rather, there are constraints in place ensuring that the protein is produced at specific times and places, and that the process of production has a clear order.

To instantiate my claims, consider the example of so-called "constitutive" or "housekeeping" genes. (See Zhu et al. 2008.) Such genes are required for the maintenance of basic cellular functions and they are always expressed, regardless of environmental conditions. Their constant expression, even under pathophysiological conditions, makes them a compelling example of a mechanism that is strongly buffered from the environment. Note, however, that the buffering in this instantiation of the mechanism of protein synthesis is not a matter of being detached from its surroundings. Rather, it is concerned with an elaborate machinery, which, nevertheless, ensures a determined responsiveness to environmental conditions. Specifically, in the group of "constitutive" genes, the levels of expression are fine-tuned to meet metabolic requirements in various tissues.

I argue that the token causal processes under consideration can be considered as instances of mechanisms in the thick sense. When a causal process gets embedded in the structure that ensures that its end product is produced in a regular and stable manner, we can reasonably claim that the emergence of a new mechanism has occurred. For this to

happen, as we have seen, it is not enough to focus on the features of the causal process itself (the steps leading from a DNA sequence to the resulting protein product), but also on the relation concerned with its proximate environment. In the case at hand, this relates to its embedding into regulatory networks already in place within a cell. I argue that the token causal processes under consideration can be considered as instances of mechanisms in the thick sense. When a causal process gets embedded in the structure that ensures that its end product is produced in a regular and stable manner, we can reasonably claim that the emergence of a new mechanism has occurred. For this to happen, as we have seen, it is not enough to focus on the features of the causal process itself (the steps leading from a DNA sequence to the resulting protein product), but also on the relation concerned with its proximate environment. In the case at hand, this relates to its embedding into regulatory networks already in place within a cell.

Let me now go back to the way in which Garson's account in question characterizes biological mechanisms in the *thick* sense – namely, as composed of different component parts, which, when aligned properly, bring about a determined phenomenon with a certain degree of regularity. I claim that it is not only a proper alignment of component parts that is important, but also the environment in which the component parts and their interactions occur. I clarify this point below.

The main underlying idea is that mechanisms, in order to produce the corresponding phenomenon in a regular manner, need to be buffered from the environment in which they are located. A similar idea is sketched by Franklin-Hall (2008), where mechanisms are characterized as a specific kind of causal structure. These structures are buffered from the context in which they occur and, additionally, can be partitioned into clearly localizable component parts responsible for the related effects. In my view, the most telling claim in the above characterization is one pointing out that mechanisms need to be appropriately independent of or buffered from other causal sequences. In my assessment, such a feature displays an immediate analogy with artifact mechanisms. To clarify further the above set of points, let me now examine the way in which artifact mechanisms can be developed by ensuring that the causal processes they contain are buffered from external influences in interesting ways.[7]

4. Artifact Mechanisms under Construction

To get a more general idea, let me start with something that would most likely be considered as a good example of a causal process, but not as a mechanism in the thick Garsonian sense (it would, however, instantiate a mechanism in his intended general, *thin*, sense). Imagine hitting a cue ball in pool and thereby breaking the rack. This process (or, roughly, a chain of interacting entities) is in no way buffered from other potential causal chains of events. Someone could push the player, and therefore influence this causal event, or if the chalk was not administered onto the cue stick, this omission could raise the chance of a miscue, among other things.

[7] I have singled out at this point Franklin-Hall's characterization of the mechanism notion, given the center stage she gives to the feature of buffering with respect to the other main features, specifically and amply debated across the New Mechanist literature, most notably, the features of organization, productive continuity, and regularity. (For their current assessment, especially, in relation to other causal structures referred to in the life sciences, such as causal pathways, see, e.g., Ross 2021.)

Consider now something that is much closer to a mechanism in the prominent biological thick sense – ball, for example, hanging from a pendulum, suspended from a pivot. In the pendulum, the ball can swing back and forth in precise intervals, i.e., it acts as a mechanism for regulating how fast the ball can move. Still, the pendulum is not very insulated from external causal influences: one can push the ball with great force, thereby causing fast oscillations, or stop them altogether. The fact that the pendulum swings back and forth in precise intervals has served as a basis for building it into a clock mechanism. In that mechanism, the movements are connected to the gears that turn the hands of the clock and are buffered from external influences by being placed in a box. Now, since the pendulum in a clock eventually slows down and stops because of friction, old clocks had to be set in motion again, either by moving the pendulum by hand (which means that this part of the clock was not entirely buffered from outside influences) or by winding it up (by an additional mechanism ensuring that the force needed for the movement of the pendulum is distributed at a constant rate). But, there is an important obstacle preventing the pendulum clock from being buffered from external causal influences to a more substantial degree – namely, they must remain stationary. Any motion or acceleration will affect the motion of the pendulum. For that reason, the corresponding next improvement of the clock mechanism consists in the invention of the spiral balance spring, controlling the oscillating speed and making mobile (wrist) watches possible.

I must emphasize here that the buffering in question is not an all-or-nothing matter. For example, a well-buffered clock mechanism can still be perturbed if we crush it with a hammer. At this point, the mechanism no longer exists. Thus, the relevant type of buffering refers to the cases in which the mechanism is engaging in its typical behavior. If an external influence is that extreme to stop the mechanism from its behavior, this is not the relevant type of buffering that I am interested in when arguing for an extended characterization of the mechanism notion with respect to the standard features highlighted by the New Mechanist approach (see also note 5). In that case, we can always envisage that any thick mechanism can be destroyed by some external event. Accordingly, at this point, I distinguish between the *buffering feature* in the case of a mere perturbation with regard to the mechanism's regular or standard operating and, on the contrary, the buffering from the destruction scenario. The latter kind of buffering is simply too demanding on any mechanism in the thick sense and, thus, can be taken into consideration in examining the malfunctioning or dysfunctioning states only to a limited extent. (For the current debate on the closely related issues pertaining to biological and biomedical functions, see Hundertmark and Bos 2024.)

Without going into further detail about the advances in a clock mechanism, the point has been about reinforcing the view according to which mechanisms need to be buffered from external influences by showing how improvements of the clock mechanism followed exactly that constraint. From the above description of the clock mechanism, I can single out some basic features of mechanisms in the thick sense. Artifact mechanisms are designed to be stable and behave regularly. My claim is that the same holds for biological mechanisms.

Before moving on to the next section, let me spell out once more that in describing the case study from the preceding section, intimately related to the mechanism of protein synthesis, I am referring to tokens of causal processes. As already pointed out in Section 3,

the evolution of a token causal process, which may end up as a token-mechanism in the thick sense co-opted by the extant mechanism of protein synthesis, proceeds via its reproduction with certain alterations through different generations of containing individual organisms. The analogy with the clock mechanism relates to the idea that a causal process remains unaffected by external influences. For a clock, this is achieved by isolating it from environmental disturbances, and adding a mechanism that ensures that it keeps running. In biological systems, such as genes, physical insulation isn't possible in that fashion. Instead, other processes ensure that the causal sequence in question remains undisturbed by integrating it into existing regulatory networks. In other words, the new process becomes part of a larger, established mechanism. Since this larger mechanism is maintained by natural selection, the newly integrated causal process is also preserved and continues to function reliably. The boundaries of this new mechanism, such as protein synthesis, are defined when a genomic sequence – previously altered randomly by mutation – becomes stabilized through natural selection. This stable genomic feature leads to a well-organized process (see here again note 1): DNA is transcribed into RNA, which is, then, translated into a new protein product. (Bear in mind that this explanation simplifies the process by omitting intermediate steps for the sake of clarity.)

5. The Upshot

Let me take stock and revisit the main features of prominently biological mechanisms or the Garsonian mechanisms in the thick sense. They are indeed historical entities in the sense that they obviously arise through evolutionary processes, and this claim represents a common ground of the New Mechanist approach, we usually do not need to rely on the thick mechanisms' historical character in the above sense when establishing whether or not something is a thick biological mechanism. However, that approach might be questioned through the following two points in Garson's "core argument": (1) the thought experiment with the "double" (Garson 2023), according to which our exact "spontaneously-generated molecule-for-molecule" twin individual organism – thus, the organism without past in any relevant sense – wouldn't have mechanisms like we standardly do – at least, so goes the etiological argument. For instance, even though our double is equal to us in every single detail, except for not having the right kind of history, its 'heart's pumping blood' is not a mechanism, as it is, on the contrary, in us mammals, for example. Thus, in this *minimal sense*, the historical dimension is needed in establishing whether something is a prominently biological mechanism.

There is yet another group of cases in which we are forced to rely on the evolutionary past in qualifying whether something is a mechanism and, in addition, what function it performs. I will refer to it as (2) the *auxiliary* historical sense: this point is about the so-called "epistemic consequence" of the "core argument". It claims that we need to make robust recourse to evolutionary history in resolving certain biological disputes, one of these disputes being concerned with the proper function of zebra stripes, for instance. (Namely, the dispute between the hypothesis saying that the function is camouflaging against predators versus the other one saying that the function is the selected activity of biting flies deterring. In brief, the latter couldn't have been confirmed without taking into account data about the trait's evolutionary past.) That's why, apart from the reasons offered by the "core argument"

itself, the historical character of prominently biological mechanisms cannot be completely ignored by the New Mechanist approach to the mechanism notion.

In my view, however, both above points can be rather easily accommodated within the approach under consideration. We can most frequently identify thick mechanisms and their functions *ahistorically*, as shown previously. Now, in those rare cases, as instantiated by points (1) and (2), the task of mechanism-function identification is not radically different from the standard ahistorical one, because the historical dimension impacts it merely marginally. On my account, a more promising way for a less marginal historical dimension in Garson's overall account is secured through the function notion, which, if historically grounded, may also influence the historical character of a mechanism in the thick sense. In fact, this might be considered as the main thread in the "core argument" (furthermore, even the normativity of prominently biological thick mechanisms is "parasitical" on the normativity of biological functions (see Garson 2023, Section 3)). But, that basically depends on the function notion we assume in the argument. Instead, we may opt for an ahistorical function notion, such as a determined version of a causal role function. In that case, if we stick to the "core argument", the historical character of biological mechanisms wouldn't be secured anymore.

To further illustrate difficulties in fleshing out quite an obvious idea of "putting history back into mechanisms", consider Garson's own examples for the class of *historical entities* to which thick biological mechanisms belong. Those examples are: igneous rocks, volcanic mountains and original Van Gogh paintings (see Garson 2023, 921-922). Garson does not distinguish between these examples, that is, they all can be equally grouped as entities with *historical essences*, as opposed to ahistorical entities, such as, most notably, gold and water. But, I would like to point out important differences between the first two examples and the third one. There are certain features (mineralogical, chemical, etc.) that igneous rocks and volcanic mountains possess, which allow us to classify them as such without invoking (geological) historical narratives in the same manner and extent as historical reconstructions concerned with the original Van Gogh paintings (and supposedly, even more so for the forged ones). Finally, the latter example is much closer to a causal trajectory, as also claimed by Garson's account, than to a thick mechanism more properly. Still, it equally has a historical essence, as prominently biological thick mechanisms do, too. As already pointed out above, when describing the *de novo* case study of gene emergence, there is a *continuum* and, sometimes, even a *progression* of different causal structures, which may end up as new token-mechanisms in the Garsonian thick sense. Thus, the situation is not a dichotomous one, as suggested by Garson's corresponding account. As a result, these recent scientific findings also suggest how to 'put history back into mechanisms', here in the paradigmatic mechanism of protein synthesis. Before concluding on that final point, let me sum up my view on thick biological mechanisms.

I have opted for a widening of Garson's current characterization of mechanisms in the following way: mechanisms in the thick sense are composed of different component parts, which, when they are aligned properly, *and are appropriately buffered from external influences*, bring about the corresponding phenomenon with a high degree of regularity (my additional characterizations are in italics). The buffering feature is meant exactly to ensure the point highlighted above of the high degree of regularity.

My main examples and considerations support the claim that in order for something to be a mechanism, it needs to have a proper function. I would argue, however, that in order to talk about functions, there already needs to be a mechanism in place ensuring a regular production of the functional outcome, as seen in the *de novo* explanatory model. If the mechanism is not already in place, then it is not possible to distinguish function from an accidental benefit.[8] Without entering into the debate on the correct notion of biological functions, I will just point out that I take the notion of a function, which is based on the notion of design (see in particular Kitcher 1993), to be suitable for characterization of mechanisms in the thick sense. The upshot of endorsing such a notion is that it applies both to biological and engineering contexts. This, in turn, allows one to ascribe functions and identify mechanisms in synthetic biology and bioengineering, as specifically argued in Section 4 of the present paper.

Finally, my analysis also emphasizes the importance of accidental causal processes occurring sporadically across the genome during the protein synthesis mechanism's normal operation. As seen, some of them result in interesting new mechanistic arrangements. Now, examining how some unstructured causal trajectories can incrementally, depending on the context, acquire more structured and embedded characteristics and, then, occasionally lead to mechanistic arrangements, reveals that they also may be considered as historical entities. That is, in cases of such diverse 'in between thin and thick' biological mechanisms, examining their gradational progression is important for establishing their genetic relevance. To put it differently, not all causal processes, in the sense of mere causal trajectories occurring in the genome, are biologically equally significant. In some cases, certain accidental arrangements can sequentially lead to relevant outcomes, if they occur in the right context. Then, in the described overall progression of events according to the *de novo* gene evolution explanatory models, it might seem that in the case of thin mechanisms, as well, we need to assess them *historically* on nearly the same grounds, i.e., the "core argument" and its "epistemic consequence", as for the Garsonian thick mechanism.[9]

Note that my use of the term 'thin mechanism' already suggests that I do not equate thin mechanisms with just any causal trajectory, as Garson does. Thin mechanisms, while

[8]Note that my proposal is compatible with, but not limited to, etiological accounts of functions. The etiological accounts hold that function ascriptions are condensed causal explanations for why a determined trait is there. But, in order for the explanation to be fruitful, not just any causal process leading up to a (functional) outcome can be invoked, but rather, the one that has the features of a mechanism in the thick sense, which I have listed above. In such a case, natural selection acts to ensure the buffering of a causal process leading up to the functional outcome. In the case of an artifact machine, conscious designers are taking the role in question.

[9]There is potentially an additional ground for assessing thin and 'in between thin and thick' mechanisms as historical entities or entities having historical essences, as claimed by Garson's account. Thus, as emphasized in Section 3, the first step in the overall *de novo* progression is, in fact, an accidental output of the mechanism of protein synthesis already in place in a cell. More precisely, the resulting non-thick mechanism is the output of the protein synthesis mechanism's "transcriptional noise". Other steps in the progression are constrained in turn by related extant biological entities, as illustrated in Figure 1. Now, this point raises the issue regarding the normative dimension of such a mechanistic operation. Namely, on the one hand, this operation is malfunctioning; but, on the other, if the protein synthesis mechanism's operation were perfect, then we wouldn't have functional novelties, at least, in the case of *de novo* progression. In my view, a further analysis of this significant issue in the functions debate goes beyond the scope of this paper.

still relatively loose, nevertheless indicate a determined degree of organization, even if a rather low one. They are interesting because of their potential to lead to evolutionarily significant outcomes. However, my aim here has not been to enter into the specifics of the debate on various kinds of mechanisms in biology, but to advance a response to a potential objection to my view on the thick, biologically prominent mechanisms and their relationship to thin mechanisms – namely, the objection mentioned by Garson (2023) against taking the machine analogy too seriously, because it fails to capture the dynamic, process-like nature of biological mechanisms. Now, I agree with the view that if there are only thick, machine-like mechanisms in biology, then a lot of interesting phenomena and causal processes would be left out of the picture.

The reported scientific case study concerned with the emergence of new genes nicely mechanisms in the thick sense. Thus, there is room for a more detailed analysis of the kinds of causal processes or structures taking place, for example, in the human genome. But, this does not mean that the kinds of causal processes under consideration cannot be best characterized as causal processes akin to those going on in the typical machines.

Acknowledgements

This paper is an output of the research project "Theoretical Underpinnings of Molecular Biology" (ThUMB), funded by the *Croatian Science Foundation*, project grant number: HRZZ-IP-2018-01-3378. I would also like to acknowledge the support from the University of Rijeka (projects "Classification in Biology and Medicine", uniri-human-18-265, and "Functions in Biology and Medicine", uniri-iskusni-human-23-227). In particular, many insightful and constructive discussions with Zdenka Brzović and Philip Kitcher are gratefully acknowledged.

References

Abrusán, G. Integration of new genes into cellular networks, and their structural maturation. *Genetics*, 195(4), 1407–17. https://doi.org/10.1534/genetics.113.152256

Carvunis, A., et al. (2012). Proto-Genes and de novo gene birth. *Nature*, 487 (7407), 370–74. https://doi.org/10.1038/nature11184

Clark, M. B., Amaral, P. P., Schlesinger, F. J., Dinger, M. E., Taft, R. J., Rinn, J. L., Ponting, C. P., Stadler, P. F., Morris, K. V., Morillon, A., Rozowsky, J. S., Gerstein, M. B., Wahlestedt, C., Hayashizaki, Y., Carninci, P., Gingeras, T. R., & Mattick, J. S. (2011). The reality of pervasive transcription. *PLoS Biology*, 9(7), e1000625. https://doi.org/10.1371/journal.pbio.1000625

Craver, C. F., & Darden, L. (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*. Chicago, IL: University of Chicago Press.

Craver, C., & Tabery, J. (2023). Mechanisms in science. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta & Uri Nodelman (eds.). Retrieved September 6, 2025, from https://plato.stanford.edu/archives/fall2023/entries/science-mechanisms/

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *American Biology Teacher*, 35(3), 125–129. https://doi.org/10.2307/4444260

Franklin-Hall, L. R. (2008). *From a Microbiological Point of View* [Unpublished doctoral dissertation, Columbia University].

Garson, J. (2023). Putting history back into mechanisms. *The British Journal for the Philosophy of Science*, 74(4), 921–940.

Hundertmark, F., & van den Bos, M (2024). Biological functions and dysfunctions: A selected dispositions approach. *Biology & Philosophy* 39 (2), 8.

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E., Wills, M. R., & Weissman, J. S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Reports (Cambridge)*, 8(5),1365–1379. https://doi.org/10.1016/j.celrep.2014.07.045

Kitcher, P. (1993). Function and design. *Midwest Studies in Philosophy* 18(1), 379–397.

Krickel, B. (2018). *The Mechanical World: The Metaphysical Commitments of the New Mechanistic Approach*. Cham: Springer Verlag.

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

McLysaght A., & Guerzoni D. (2015). New genes from non-coding sequence: The role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 370 (1678), 20140332.

Peng J., & Zhao, L. (2024). The origin and structural evolution of de novo genes in *Drosophila*. *Nature Communications*, 15, 810. doi.org/10.1038/s41467-024-45028-1.

Ross L. N. (2021). Causal concepts in biology: How pathways differ from mechanisms and why it matters. *British Journal for the Philosophy of Science*, 72(1), 131–158.

Schmitz J. F., & Bornberg-Bauer E. (2017). Fact or fiction: Updates on how protein-coding genes might emerge *de novo* from previously non-coding DNA. *F1000 Research*, 19(6), 57. https://doi.org/10.12688/f1000research.10079.1

Tautz, D. (2014). The discovery of de novo gene evolution. *Perspectives in Biology and Medicine*, 57(1), 149–161. https://doi.org/10.1353/pbm.2014.0006

Tautz, D., & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews. Genetics*, 12(10), 692–702. https://doi.org/10.1038/nrg3053

Van Oss S. B., & Carvunis A. R. (2019). De novo gene birth. *PLoS Genetics*, 15(5), e1008160. https://doi.org/10.1371/journal.pgen.1008160

Weisman, C. M. (2022). The origins and functions of de novo genes: Against all odds? *Journal of Molecular Evolution*, 90, 244–257. https://doi.org/10.1007/s00239-022-10055-3

Zhu, J., He, F., Hu, S., & Yu, J. (2008). On the nature of human housekeeping genes. *Trends in Genetics*, 24(10), 481-484. https://doi.org/10.1016/j.tig.2008.08.004

Balkan Journal of Philosophy (2025), vol. 17, issue 2, 124–137

RECONCILING AI AND PATIENT AUTONOMY: A CONSEQUENTIALIST FRAMEWORK FOR INFORMED CONSENT IN HEALTHCARE

Joby Varghese*

Abstract: One of the most significant ethical concerns in the context of the use of AI in healthcare is the patient's autonomy and informed consent. The problem of autonomy emerges from the concern that black box AI could endanger patient autonomy because its use in healthcare is incompatible with informed consent practice in healthcare. In this paper, I argue that AI's opaque workings can be reconciled with morally justified healthcare practices through a coherent ethical framework based on consequentialism, and AI can preserve the autonomy of patients by guaranteeing transparency, disclosure and probabilistic assessment during the treatment process. Further, I will contend that although AI's decision-making process remains opaque in terms of the internal workings of its algorithms, this lack of transparency does not necessarily undermine the ethical principle of informed consent if we employ "*The rule of transparency and probabilistic assessment*" grounded on a consequentialist framework.

Keywords: AI in healthcare; Black Box problem; Patient's Autonomy; Informed consent; Rule utilitarianism; Ethics of AI.

1. Introduction

The emergence of AI trends signals a paradigm shift in several industries and presents unprecedented growth opportunities. Recently, AI has proven not only to be essential to innovation in a variety of industries such as manufacturing, healthcare, banking, and logistics, but also it has simplified procedures and redefined norms. Artificial Intelligence (AI) has become a ground-breaking technology in the last fifteen years, addressing and resolving societal issues and promoting economic growth. AI has the potential to significantly improve production and efficiency while addressing global problems, including poverty, healthcare inequities, and climate change (Floridi & Cowls, 2022; Mittelstadt et al., 2016). AI is used in many different industries.

AI is having a revolutionary effect on healthcare. By utilizing patient data and genetic information, it facilitates individualized diagnostics and treatments, refining results and maximizing medical research (Topol, 2019). AI tools, for example, help determine the best course of treatment and predict patient outcomes (Sullivan & Schweikart, 2019). But these transformative powers of AI technologies do bring up moral questions. One of the most significant ethical concerns is regarding the patient's autonomy when AI is used in healthcare. Since AI is generally considered to be a "black-box" that obscures decision-making processes, critics contend that it compromises informed consent (Floridi & Cowls, 2022; Mittelstadt et al., 2016).

*Department of Humanities & Social Sciences, Indian Institute of Technology Jammu; joby.varghese@iitjammu.ac.in; ORCID ID: https://orcid.org/0000-0002-3108-3406

The primary aim of this paper is to argue that AI's opaque workings can be reconciled with morally justified healthcare practices through a coherent ethical framework based on consequentialism, and AI can preserve the autonomy of patients by guaranteeing transparency and patient participation during the treatment process. In this essay, I will contend that, despite the apparent incompatibility between informed consent and the use of AI in healthcare, the black box problem can be pragmatically confronted by framing and applying a consequentialist ethical framework.

2. Ethical issues in the application of AI in healthcare

The application of AI in healthcare brings forth various ethical concerns regarding fairness, transparency, autonomy, explainability and others. These concerns should be addressed in an ethically plausible manner in order to ensure equity and trust in AI-driven healthcare practices. For instance, AI systems in healthcare often operate as "black-box" models, making their decision-making processes opaque to both clinicians and patients. This lack of transparency undermines trust and accountability, particularly in high-stakes situations like diagnostics or treatment recommendations (Mittelstadt et al., 2016; Sullivan & Schweikart, 2019). Similarly, AI algorithms that are biased due to various factors can disadvantage underrepresented groups and prolong healthcare inequities. For example, racial biases in training data have exacerbated healthcare inequities by producing less accurate predictions for minority groups (Obermeyer et al., 2019). In this regard, adopting fairness-aware machine learning techniques, auditing algorithms for bias, and utilizing a variety of datasets are all necessary to ensure fairness (Holmes et al., 2019; Dilsizian & Siegel, 2014). There is also the problem of explainability. Clinical and patient decision-making is hampered by the inability to explain AI judgments adequately. Most importantly, ethical questions regarding accountability and informed consent often emerge because of the lack of explicability. Although explainable AI (XAI) techniques are being developed, including interpretable models and visualization tools (Tonekaboni et al., 2019; Floridi & Cowls, 2022) to address the problem of explainability, there are still concerns, and one of the most significant concerns is patient autonomy.

The application of AI in healthcare raises concerns about patient autonomy, especially when those systems take precedence over or significantly impact human judgment. Since patients and physicians may find it difficult to completely comprehend AI suggestions, the black-box nature of AI makes informed consent even more challenging (Sullivan & Schweikart, 2019; Topol, 2019). In order to retain autonomy while utilizing AI's advantages, this study proposes a rule-utilitarian approach that places an emphasis on transparency, disclosure and probabilistic assessment.

3. Autonomy as a Normative Concept

Autonomy is a normative concept in moral and ethical philosophy that describes people's ability to govern themselves in conformity with reasonable norms. It symbolizes the capacity to make one's own decisions based on morality and reason, and it is fundamental to concepts like freedom, accountability, and moral agency. Philosophers have approached autonomy in a number of ways, assigning various values. Let us briefly discuss how the concept of autonomy is understood in a consequentialist theory, which is utilitarianism.

3.1. Autonomy in Utilitarianism

Utilitarianism is a moral theory that evaluates actions based on their consequences, aiming to maximize overall happiness and minimize suffering. It is considered that the best action is the one that produces the greatest good for the greatest number. There are two different forms of utilitarianism, which are act utilitarianism and rule utilitarianism. Act utilitarianism and rule utilitarianism have distinct perspectives on autonomy, which is defined as the ability to regulate oneself and make logical decisions. Both theories are part of the larger utilitarian tradition, but because they place different emphasis on individual actions and broad principles, they regard autonomy differently. These two ethical frameworks are examined below, with an emphasis on the philosophical nuances and pragmatic usages of autonomy.

3.1.1. Autonomy in Act Utilitarianism

Act utilitarianism assesses the morality of any action according to that particular action's potential to maximize utility or happiness. Autonomy is instrumental according to act utilitarianism, valued mostly for its capacity to bring forth positive results. Respecting autonomy is maintained when doing so increases overall utility or happiness. However, it may be justified when sacrificing autonomy results in increased satisfaction or less suffering. The principle of utility takes precedence over autonomy due to its conditional respect (Mill, 1863). In the healthcare industry, for example, an act utilitarian could defend disregarding a patient's autonomy to decline life-saving treatment if doing so would improve the overall outcomes, including saving the patient's life and maintaining their well-being in the future (Singer, 2011). According to this viewpoint, autonomy is considered as just one element without any inherent value among many others that determine utility.

3.1.2. Autonomy in Rule Utilitarianism

Rule utilitarianism, on the other hand, bases its assessment of an action's morality on its compliance with rules that, in general, will produce optimal happiness or utility. Given its long-term benefits in fostering trust, cooperation, and dignity of moral agents, this framework provides a more steady and systematized approach to autonomy. Rule utilitarianism aims to strike a balance between individual rights and the welfare of the group by giving priority to regulations that respect autonomy (Hooker, 2000). Respecting autonomy in healthcare scenarios, for example, by getting informed permission, can increase public confidence in healthcare systems and improve the overall promotion of health. A rule that always respects autonomy by upholding moral principles and lowering the fear of coercion, even if it might not maximize benefit in a particular situation, contributes to increased long-term utility (Beauchamp & Childress, 2001). The idea that respecting autonomy results in just and predictable systems is consistent with rule utilitarianism's larger emphasis on society's welfare. For instance, rules that promote individual judgment in democratic government or individual liberties in moral quandaries create spaces where people feel respected and empowered. This respect for autonomy indirectly increases societal utility through mutual respect and trust (Mill, 1863).

3.2. Flexibility vs. Consistency: Autonomy in Utilitarian Ethics

The way in which act and rule utilitarianism view autonomy is very different. The situational flexibility of act utilitarianism permits complex reactions to specific situations, but it also runs the risk of eroding consistency and trust. However, rule utilitarianism's commitment to specific moral rules provides a solid foundation for maintaining autonomy, especially in institutional or social contexts where consistency and fairness are essential. From a philosophical perspective, autonomy is more consistent with rule utilitarianism's methodical approach. Most importantly, this theory endorses the intrinsic value of self-determination and autonomy even in such contexts where upholding these concepts might compromise the immediate utility. Yet, they are preserved because preserving them would produce more utility in the long run rather than not preserving them. Act utilitarianism, in contrast, views autonomy as a conditional element that is only worthwhile when it contributes to the objective of promoting happiness (Singer, 2011). Practically speaking, upholding autonomy generally promotes moral decision-making in a variety of fields, such as healthcare, law, and education. It ensures that people feel valued and empowered by fostering cooperation and trust. In what follows, I will discuss the problems regarding autonomy and informed consent, which pop up because of the opacity of AI decision-making in the context of healthcare.

4. The problem of autonomy and the black box issue in the application of AI in Healthcare

In the context of healthcare, autonomy refers to a person's ability to make well-informed decisions regarding their medical care, supported by a thorough comprehension of the rationale behind such decisions. However, patients – and occasionally even medical professionals – may not have adequate knowledge when AI systems are involved in diagnosing a disease or recommending a treatment. Without clear explanations of the reasoning behind AI-generated outcomes, healthcare providers may struggle to justify treatments, and patients may feel disempowered in making informed medical choices. This challenge, often referred to as the black-box problem, raises serious ethical concerns regarding accountability, trust, and informed consent in AI-driven healthcare. From a moral perspective, the problem of the "black box", which refers to the opacity of AI decision-making, is closely related to the autonomy issue in the application of AI in healthcare. The black box assertion draws attention to the difficulty of not being able to completely comprehend or explain how AI systems generate particular results or reach particular conclusions. There are serious ethical issues with this opacity, particularly with regard to patient autonomy.

In traditional medical practice, a patient can ask the treating doctor for details about why the doctor has made a diagnosis or suggested a course of therapy, and the doctor is able to provide an explanation of their thought process, diagnostic rationale, and supporting data. A fundamental tenet of medical ethics, which is informed consent, is preserved by this transparency. In contrast, AI operates on vast amounts of data and complex algorithms that may not be easily interpretable. AI systems often provide results without detailed explanations of how those results were derived, leaving both patients and healthcare providers in the dark regarding the rationale behind AI-driven decisions (Morley et al., 2020). This lack of transparency can undermine the trust necessary for informed consent.

The black box problem becomes more problematic because, unlike a human doctor who can explain their rationale, AI's decision-making process is not always transparent to people who depend on its results. Even the developer's AI systems may sometimes fail to fully comprehend how specific outputs are produced, especially in deep learning models that learn from data patterns rather than predefined rules (Burrell, 2016). This epistemic ambiguity is extremely challenging in such contexts where patients will have to make well-informed decisions about their treatments. Consequently, patients' autonomy is undermined since they are unable to completely weigh the risks, advantages, and available options when they are not in a position to understand how an AI system arrives at a diagnosis or therapy decision.

The philosophical and ethical concerns raised by the black box problem can be illustrated through the following argument:

(1) A morally justified treatment of patients necessitates respect for the patient's autonomy (Beauchamp & Childress, 2001).

(2) Respect for the patient's autonomy entails informed consent (Faden & Beauchamp, 1986).

(3) Treatments based on the decisions of AI are opaque and thus constitute black boxes (Morley et al., 2020).

(4) Black boxes are incompatible with informed consent (Burrell, 2016).

(5) Therefore, morally justified treatment of patients in healthcare cannot be based on black box AI.

This argument suggests that since AI systems cannot provide the transparency needed for informed consent, their use in healthcare may violate patient autonomy. Consequently, any medical treatment involving AI that fails to respect autonomy would be morally problematic. In the following section, I shall adopt a rule utilitarian ethical framework and make an attempt to show that premise 4, *black boxes are incompatible with informed consent*, is faulty, and hence, the above-given argument will not hold. But before we discuss the rule utilitarian framework, it is necessary to have a glance at the notion of 'informed consent' in traditional medical practice.

5. Informed Consent in Healthcare

Informed consent in healthcare is a foundational principle of medical ethics and law, ensuring that patients are fully informed about and they understand the potential risks, benefits, and alternatives before agreeing to a particular medical procedure or treatment. It highlights people's right to make their own healthcare decisions after receiving all pertinent information from medical specialists. To guarantee informed consent, the patient must be made fully aware of a number of factors. Before gaining informed consent, the patient should be clearly informed of the following conditions and information.

(i) *Nature of the Procedure or Treatment:* The patient must be informed about the specifics of the procedure or treatment, including what it entails and the steps involved. This is necessary to ensure that the patient is making a decision based on a

clear understanding of what is going to happen (Faden & Beauchamp, 1986; Miller & Brody, 2003).

(ii) *Risks and Benefits:* Patients must be made aware of both the potential benefits and risks associated with the treatment or procedure. This includes common side effects, the likelihood of success, and any potential complications. The disclosure should be comprehensive, allowing the patient to make an informed decision based on an understanding of the potential outcomes (Mackenzie et al., 2014; O'Neill, 2002).

(iii) *Alternatives:* If there are reasonable alternatives to the suggested course of treatment, then they should be explained to the patient. The patient should understand the relative risks and benefits of each of these options, which may include various treatments, surgeries, or strategies. To guarantee that patients have a wide range of options, it is essential to comprehend the alternatives (Beauchamp & Childress, 2001; Miller & Brody, 2003).

(iv) *Consequences of Non-Treatment:* The possible repercussions of not receiving the suggested treatment must be explained to patients. This could involve the course of the illness or condition and the possibility of adverse consequences in the event that treatment is refused (Beauchamp & Childress, 2001).

(v) *Capacity to Understand:* The patient must be able to comprehend the information being given to them. The patient must be cognitively capable in order to digest and understand the specifics of the treatment. Special efforts may be necessary in certain situations to assure understanding due to language barriers or cognitive impairments (Mackenzie et al., 2014; O'Neill, 2002).

(vi) *Voluntariness:* The choice must be freely made – free from undue influence or force. This implies that the patient is free to think through their alternatives, ask questions, and make a decision based on their own values and preferences (Beauchamp & Childress, 2001; Faden & Beauchamp, 1986).

(vii) *Documentation:* To serve as a record of the procedure and guarantee that the patient's consent has been appropriately acquired, informed consent should ideally be recorded, usually with a signed consent form. However, in other situations, verbal consent can also be acceptable (Faden & Beauchamp, 1986).

Giving informed consent is a continuous process rather than a one-time occurrence. If a patient's condition changes or new information becomes available, the patient might need to go through the consent procedure again. Ultimately, the objective is to respect patient autonomy and let people make choices that are consistent with their values and knowledge of their healthcare options. Moreover, the right to self-determination is the foundation of informed consent. This implies that the treating physician should provide patients with all the information that is available so that they can make an informed choice about the suggested course of therapy.

The concept of informed consent states that a patient should be made aware of the potential advantages and disadvantages of specific treatments before voluntarily choosing the suggested course of action. Not all information regarding the diagnosis and potential advantages and disadvantages is easily accessible to human doctors, even in the case of

traditional medical practices. Nonetheless, the patient's autonomy is sufficiently upheld because the knowledge at hand is thought to be adequate for informed consent for various reasons.

Firstly, a reasonable disclosure of important information, such as the nature of the therapy, its risks and benefits, available alternatives, and the repercussions of not receiving treatment, is sufficient to meet the principle of informed consent rather than complete knowledge. As long as these elements are met, patient autonomy is respected. Secondly, even in traditional medicine, diagnoses and treatments rely on probabilities and clinical judgment rather than absolute certainty. This uncertainty does not undermine autonomy as long as patients receive enough information to weigh their options rationally. Thirdly, there is a normative reason. Doctors are ethically and legally bound to provide the most accurate and relevant information possible based on current medical knowledge. The assumption is that medical institutions, research, and professional oversight ensure that patients receive sufficiently reliable information. In other words, whether informed consent is sought in the context of diagnosis and treatment by a human physician or an artificial intelligence system, it is sufficient that whatever available information about the treatment procedure, risks and benefits, alternative treatment options are provided and other guidelines, such as obtaining voluntary consent and documentation of informed consent, are duly followed. Now the question is: how does one make sure that all available information required for obtaining informed consent is clearly provided and other procedures are duly followed in the context of the application of AI in healthcare? This question can be addressed by asking a few other questions, which are framed based on rule utilitarian theory, and addressing them. But before we discuss the question, let us quickly go over two recent works on integrating Artificial Intelligence (AI) into healthcare and the concerns regarding patient consent.

Cohen (2024) examines the obligations of doctors to disclose AI involvement in medical decisions. Given the complexity and opacity of AI systems, he investigates whether further disclosures beyond the conventional norms for informed consent are necessary when using AI. He argues that because of the complexity and opacity of AI, additional disclosures are required in addition to traditional informed consent. In a similar vein, Pruski (2024) contends that consent mechanisms already in use are adequate for AI-enhanced treatment. Pruski argues, with reference to the National Health Service in the United Kingdom, that AI does not provide essentially new ethical issues necessitating modifications to existing informed consent practices. However, Pruski's theory underestimates the impact of AI's opacity. He assumes that AI can be treated like traditional medical technologies, but unlike human decision-making, AI systems do not provide intuitive explanations. Patients may not just want to know what the AI decided, but why – something current consent models do not adequately address.

Although both Pruski and Cohen offer valuable viewpoints, a middle path might be needed here. Pruski's position seems to underestimate unique ethical challenges in the context of the use of AI in healthcare, specifically with reference to informed consent, while Cohen's need for more disclosures about the technology runs the danger of overwhelming patients and taxing physicians. A middle path where the patients are given brief explanations with the option to view more in-depth AI-related information if they so choose could be a

workable alternative. Here, the best approach would be to look for a practical solution which is normative in nature.

6. A Consequentialist Framework to resolve the Problem of Autonomy

When actions are evaluated from a moral standpoint, utilitarianism concentrates on their utility or the consequences. Rule utilitarianism suggests that "the rightness or wrongness of a particular action is a function of the correctness of the rule of which it is an instance" or that an action is right if it complies with a rule that maximizes good (Garner & Rosen, 1970). In this section, I will use the rule utilitarian framework and show that the premise 'Black boxes are incompatible with informed consent' is false, and hence, the claim that 'morally justified treatment of patients in healthcare cannot be based on black box AI' is flawed.

The transparency and comprehensibility of the information provided to patients are essential for obtaining their informed consent in the medical field. However, it is important to remember that neither all AI-processed data nor the intricacy of its analysis is always pertinent to informed consent. There are fundamental components that must be conveyed in a way that facilitates a thorough decision-making process, even though it may not be necessary for the patient to comprehend every aspect of the algorithm's operation. The necessary information about those fundamental components can be provided by formulating a rule from a rule utilitarian perspective. The rule that can be employed in the context of the use of AI in healthcare and obtaining informed consent is 'The rule of transparency and probabilistic assessment'. This rule is grounded in the principle of maximizing overall well-being, as providing the necessary information, such as the nature of the treatment, risks, benefits, alternatives, etc., and increases the likelihood that patients will make well-informed decisions that align with their values and interests. Similarly, this rule will also help maximize patient well-being by providing a clear understanding of potential outcomes, both positive and negative, based on the probabilities of success and failure.

One way to maintain the rule of transparency and probabilistic assessment is to frame a few necessary questions that AI should be able to answer in order to sufficiently preserve informed consent. These questions align with key factors used in utilitarian ethics to evaluate the moral dimensions of an action. According to utilitarian theorists, the moral worth of actions should be assessed based on the following five criteria: (i) net amount of happiness, (ii) duration, (iii) likelihood, (iv) intensity, and (v) fruitfulness (Bentham, 1879; Mill, 1863). Using these factors as guidelines, the AI should be able to respond to the following critical questions in a probabilistic manner.

(i) *Does the net benefit of the treatment outweigh the potential harms?* In the context of AI in healthcare, determining whether the net benefit of a treatment outweighs its potential harms involves assessing the outcomes based on AI-driven predictions. For example, an AI system may analyze medical imaging to identify early-stage cancer. If the treatment recommended by AI, such as chemotherapy or surgery, can significantly reduce the risk of mortality but carries side effects like hair loss or nausea, the patient must understand whether the long-term benefit, cancer remission, justifies these temporary harms. A better understanding of the benefit-risk ratio may be possible because of AI's capacity to quantify these trade-offs using extensive

datasets (Davenport & Kalakota, 2019; O'Neill, 2002). Addressing this question in a probabilistic manner will provide the information required for the nature of the treatment recommended and the potential benefits and risks.

(ii) *How long will the benefits and harms persist?* AI can also predict the duration of both the benefits and harms of a treatment. For instance, AI might predict that a particular medication will lower blood pressure for several years, but it may also highlight the long-term side effects, such as potential kidney damage. AI models can analyze long-term clinical data to offer insights into how long the treatment's effects will last and how long any side effects might persist. This ability to provide long-term projections is especially valuable in managing chronic conditions and allowing patients to make informed decisions about whether to accept treatments that have long-term consequences (Beam & Kohane, 2018; Esteva et al., 2017).

(iii) *How likely are these advantages and disadvantages to materialize?* Using predictive algorithms and historical data, AI in healthcare can offer accurate probability. For instance, an AI system might predict a 90% success rate for a heart surgery while also calculating a 5% risk of severe post-operative complications. This level of probabilistic prediction helps patients understand the chances of both positive outcomes, such as recovery or improved health, and negative outcomes, such as infection or complications. By analyzing enormous volumes of patient data, AI can determine these risks more accurately than conventional techniques, ultimately supporting better informed decision-making (Rajkomar et al., 2019; Meskó et al., 2017).

(iv) *How intense are the positive and negative outcomes of the treatment?* The intensity of treatment outcomes in AI-assisted healthcare is a crucial consideration. For instance, an AI system might recommend a radical surgery for treating aggressive cancer. The positive outcome (extended life expectancy) is significant, but the negative outcome (post-surgical pain, long recovery) could be intense. AI models can quantify the intensity of both sides by considering patient-specific factors such as age, comorbidities, and overall health. These personalized predictions help patients assess whether the intensity of the benefits justifies the intensity of the harms. AI's ability to personalize these predictions based on individual data allows for more accurate evaluations of treatment intensity (Topol, 2019; Obermeyer et al., 2019). Addressing this question might also give information about the consequences of non-treatment and also the available alternative treatments.

(v) *How fruitful is the recommended treatment in terms of long-term health benefits and prevention of future health issues?* The AI must evaluate the treatment's wider, long-term impacts in addition to its immediate ones in order to respond to this query. For instance, the AI might forecast the chance of remission, long-term survival, and relapse prevention in the context of cancer treatment. It might also take into account the possibility of lower future medical expenses or a better quality of life for the patient. A treatment which is so successful should ideally provide long-term advantages and reduce the risk of recurrent health issues. Effective communication of these long-term effects by the AI enables the patient to comprehend not only the treatment's immediate effects but also its potential for the future.

Asking these questions, which are based on the rule of transparency, disclosure, and probabilistic assessment, and providing answers to these questions as given above can give patients the information they need to make an informed, voluntary decision. The AI should present this information clearly and probabilistically, similar to how a human doctor would explain potential outcomes based on medical knowledge and experience. In this sense, even though the AI's decision-making process remains opaque in terms of the internal workings of its algorithms, which leads to the problem of the "black box", this lack of transparency does not necessarily undermine the ethical principle of informed consent, as long as the necessary questions are addressed effectively. This perspective suggests that the black box problem, while concerning, does not automatically invalidate the autonomy of patients in the context of AI-assisted healthcare. As long as the AI can address these critical questions probabilistically and the answers are communicated in an understandable way, informed consent can still be considered ethically valid. As a result, even with little internal process openness, AI systems may nonetheless encourage autonomy and well-informed decision-making (O'Neill, 2002).

7. Extended Problems

A question that may be asked in this context is whether informed consent genuinely prevents misunderstandings akin to the "black box" problem in the context of AI. This is a difficult question to address because of the complex nuances involved. The reason is that while the purpose of informed consent is to empower patients with enough understanding to make voluntary, autonomous choices, several inherent limitations in how informed consent works, especially in real-world clinical settings, make it poorly equipped to address the kind of opacity presented by many AI systems.

Informed consent is built on three pillars: disclosure, comprehension, and voluntariness. While AI systems may disclose probabilistic outcomes and treatment recommendations, they often fail to support genuine comprehension, especially regarding how and why a specific decision or recommendation was made. When it is said that the nature of AI is a "black box", it also means that even developers are fully unable to comprehend the algorithm's internal logic. An implication of this scenario is that while outcome-level transparency might be offered, the reasoning behind decisions still remains obscure, which may lead to possible misunderstandings or misplaced trust.

The intricacy of medical knowledge poses a persistent challenge to informed consent. Because of the use of technical jargon and a lack of health literacy, patients often find it difficult to understand the consequences of diagnosis, procedures, and therapies. Patients may become overwhelmed by the cognitive load of understanding risks, benefits, and choices associated with diagnosis, procedures, and therapies, even when explanations are provided in simple terms. This challenge, which arises in the context of traditional medical practices, can also be observed in the use of AI in healthcare. For instance, it is similarly difficult to describe how an algorithm determines a diagnosis or suggested course of treatment. It is an extension of a larger problem in which patients frequently give their consent without fully understanding traditional medical decisions. The opacity involved in an AI diagnostic or therapy prescription is, therefore, not essentially novel.

It is hardly controversial to claim that paternalism in medicine continues to limit patient autonomy. Traditional medical culture often leans toward paternalism, where healthcare

providers make decisions for patients "in their best interest." In these situations, professional judgment shapes informed consent more than patient empowerment. The reason is that the information supplied might be deliberately filtered to support the clinician's opinion, and this could restrict the scope of informed consent. This dynamic may get worse with AI. It is highly probable that since clinicians believe algorithmic recommendations are more authoritative or accurate than human judgment, they may depend too much on them. On top of that, if AI-generated judgments are presented as definitive or indisputable, the involvement of patients in collaborative decision-making may be further diminished.

Another significant issue that needs to be considered in the context of informed consent is information asymmetry. Patients typically lack the expertise to question or critique medical recommendations and hence, they typically rely heavily on the expertise of healthcare providers. This power imbalance often results in trust-based, rather than knowledge-based, consent. In other words, informed consent is, therefore, more about placing trust in a course of action than making a fully informed choice. AI exacerbates this asymmetry because both patients and many clinicians may not fully understand the workings of the system. When neither party can critically assess how a diagnosis or treatment plan was generated, informed consent becomes a ritual rather than a genuine ethical safeguard.

8. Possible Objections

One of the most reasonable objections that could be raised in the context of using a rule utilitarian framework to morally justify the application of black box AI in healthcare is why a rule utilitarian framework should be privileged rather than other consequentialist theories, such as act utilitarianism. A rule utilitarian approach is particularly effective in addressing the problem of informed consent in the use of black box AI in healthcare because it emphasizes the importance of consistent and universally beneficial rules regarding transparency and patient autonomy, which will promote the patients' overall well-being. By adhering to rules that prioritize informed consent, ensuring patients understand AI's role, risks, and limitations, rule utilitarianism promotes trust, reduces harm, and upholds patient autonomy and dignity. In contrast, act utilitarianism might not care much about the autonomy of patients or informed consent and even bypass or compromise these critical factors to maximize patients' happiness in specific treatment scenarios. Act utilitarianism might even justify withholding information or manipulating outcomes in specific cases for the perceived greater good. In that way, a rule utilitarian approach is better suited for addressing transparency and probabilistic assessment of treatment-related outcomes in the context of black box AI in healthcare because it prioritizes establishing consistent, ethical rules that maximize overall well-being. Thus, the rule utilitarian approach is superior to other consequentialist theories in addressing informed consent issues with black-box AI in healthcare because it provides a structured, reliable framework that balances technological innovation with ethical responsibility in patient care. Moreover, this approach aligns with the broader goal of maximizing societal well-being while maintaining moral integrity in healthcare practices, and by requiring explainable AI and clear communication of probabilistic treatment outcomes, rule utilitarianism fosters patient autonomy while maintaining ethical integrity and trust in AI-driven healthcare decisions.

Another objection one might raise here is the difference in the type of reasoning a human doctor makes and an AI tool uses while diagnosing a patient and recommending treatment. In the context of informed consent, it is essential that whatever available information about the treatment procedure, risks and benefits, and alternative treatment options is provided to the patient. Further, the patient may also be given an explanation of how a certain diagnosis has been made. In the case of human doctors, they often rely on causal reasoning, combining patient history, symptoms, and clinical knowledge to infer the underlying cause of a disease. On the other hand, AI tools make use of statistical reasoning for identifying patterns, correlations, and probabilities based on large datasets. AI tools excel at this, as they can process vast amounts of data quickly and identify statistical associations. However, while statistical inference can highlight correlations, it does not establish causation. This can lead to misinterpretations, especially in complex medical cases where confounding factors are present.

For accurate disease diagnosis, a combination of both approaches is ideal. AI tools can provide statistical insights and highlight potential issues, while human physicians can apply causal reasoning to interpret these findings in the context of the patient's unique situation. This collaborative approach leverages the strengths of both statistical and causal inference, improving diagnostic accuracy and patient outcomes. When combined with causal reasoning, AI tools become more reliable by distinguishing between mere associations and true disease mechanisms. Rule-based AI models incorporating causal inference can improve diagnostic accuracy, ensuring both physicians and AI tools move beyond correlations to evidence-based treatment decisions. Thus, integrating causal inference with statistical methods provides the most robust diagnostic approach in healthcare. However, the difference in the reasoning of both human doctors and AI systems and its impact on informed consent in the context of the use of black box AI requires further attention.

9. Conclusion

In this paper, I have presented an argument challenging the claim that morally justified treatment in healthcare cannot rely on black box AI. The black box problem posits that the opacity of AI systems – where the decision-making process remains unexplained – undermines informed consent and, by extension, patient autonomy. In this context, I have argued that if AI systems can answer certain key questions based on the rule of transparency, disclosure, and probabilistic assessment related to treatment outcomes, that information will be sufficient for patients to make informed, voluntary decisions. Therefore, as long as AI judgments allow patients to comprehend the ramifications of their treatment choices, they can still be consistent with the autonomy principle, even if the internal workings of the algorithms remain opaque.

In conclusion, while the opacity of AI decision-making raises concerns about transparency, it does not necessarily undermine the ethical validity of informed consent. As long as AI systems clearly present relevant information probabilistically – much like human doctors do when explaining potential outcomes – patients can still make informed, voluntary decisions. The key is ensuring that critical questions about risks, benefits, and alternatives are addressed effectively, even if the internal workings of the AI remain complex. Therefore,

despite limited transparency in AI's internal processes, these systems can still support patient autonomy and well-informed decision-making when properly designed and communicated.

Acknowledgements

This research was supported by the Indian Institute of Technology Jammu for the project '*Ethical issues concerning the design and deployment of AI systems*' under Grant No. SGT-100059. The author gratefully acknowledges this support, which made the study possible.

References

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA, 319*(13), 1317–1318.

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics* (5th ed.). Oxford University Press.

Bentham, J. (1879). *The principles of morals and legislation*. Clarendon Press.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), https://doi.org/10.1177/2053951715622512

Cohen, I. G., & Slottje, A. (2024). Artificial intelligence and the law of informed consent. In *Research handbook on health, AI and the law* (pp. 167–182). Edward Elgar Publishing.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal, 6*(2), 94–98.

Dilsizian, S. E., & Siegel, E. L. (2014). Artificial intelligence in medicine and cardiac imaging. *Current Cardiology Reports, 16*(1).

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118.

Faden, R. R., & Beauchamp, T. L. (1986). *A history and theory of informed consent*. Oxford University Press.

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, 535-545.

Garner, R. T., & Rosen, B. (1970). A Systematic Introduction to Normative Ethics and Meta-Ethics. *Journal of Symbolic Logic, 35*(3).

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education promises and implications for teaching and learning*. Center for Curriculum Redesign.

Hooker, B. (2000). *Ideal code, real world: A rule-consequentialist theory of morality*. Oxford University Press.

Mackenzie, C. (Ed.). (2014). *Vulnerability: New essays in ethics and feminist philosophy*. Oxford University Press, USA.

Meskó, B., Drobni, Z., Bényei, É., Gergely, B., & Győrffy, Z. (2017). Digital health is a cultural transformation of traditional healthcare. *mHealth, 3*(38). https://doi.org/10.21037/mhealth.2017.08.07

Mill, J. S. (1863). *Utilitarianism*. Parker, Son, and Bourn.

Miller, F. G., & Brody, H. (2003). A critique of clinical equipoise: Therapeutic misconception in the ethics of clinical trials. *Hastings Center Report, 33*(3), 19–28.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1–21.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). The ethics of AI in healthcare: A mapping review. *Social Science & Medicine, 260*, 113172.

Obermeyer, Z., Powers, B. W., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

O'Neill, O. (2002). *Autonomy and trust in bioethics*. Cambridge University Press.

Pruski, M. (2024). AI-enhanced healthcare: Not a new paradigm for informed consent. *Journal of Bioethical Inquiry, 21*(3), 475–489.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358.

Singer, P. (2011). *Practical ethics*. Cambridge university press.

Sullivan, H. R., & Schweikart, S. J. (2019). Are current tort liability doctrines adequate for addressing injury caused by AI? *AMA Journal of Ethics, 21*(2), E160–166.

Tonekaboni, S., et al. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Proceedings of Machine Learning Research*.

Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Hachette UK.

Balkan Journal of Philosophy (2025), vol. 17, issue 2, 138–145

PARADOX AND MODALITY

Blagovest S. Mollov*

Abstract: The text identifies a puzzle about modal paradoxes: that if solvable, they are pseudo-paradoxes, and if insoluble, they fail to be instructive. I illustrate the point with the indicative case of Chisholm's paradox and point to possible deflationary uses for paradoxes.

Keywords: A puzzle about paradoxes; Chisholm's modal paradox; modal tolerance; iterated modality.

1. Introduction

Paradoxes are standardly taken to be (apparently) sound arguments that lead to (apparently) absurd or false conclusions.[1] Respectively, resolving paradoxes involves either showing (a) that (contrary to appearances) the argument is not sound – is formally invalid, contains one or more false premises, or instantiates an informal fallacy; or (b) that (contrary to appearances) the conclusion is not false or absurd[2]; in both cases, with an additional explanation of the misleading appearances. In Quine's popular taxonomy (Quine, 1966), (a) and (b) are respectively labelled "falsidical" and "veridical" paradoxes. To these Quine adds a third class of paradoxes, antinomies, that cannot be (a)- or (b)-resolved and thus are said to "bring on the crises in thought ... establish[ing] that some tacit and trusted pattern of reasoning must be ... avoided and revised"; and that "can be accommodated by nothing less than a repudiation of part of our conceptual heritage" (Quine, 1966, pp. 7, 11).

In this article, I will argue that the problems I take the standard Quinean taxonomy to have are symptomatic of a deeper one. I will sketch what I call a "puzzle about modal paradoxes" and illustrate it with the case of "Chisholm's modal paradox" (CMP). Though I propose a deflationary approach, there can arguably be a stronger inclination to draw an eliminative conclusion.

2. Issues

The standard (Quinean) taxonomy can be found objectionable on several counts. It can be argued that fallacious arguments should not be considered paradoxes proper since there is nothing paradoxical in deceptive appearances of argumentative soundness. At best, a fallacious argument for a non-conventional conclusion can be classified as a pseudo-paradox but would still be of only logical or rhetorical, rather than conceptual or ontological, interest. It can be further claimed – for the same reason – that no paradox that can be (a)-resolved,

[1] From Gr. πᾰ́ρᾰ (pára), "contrary to" and δόξᾰ (dóxa), "belief, opinion".

[2] If the paradoxicality of the conclusion consists in its contradictoriness, adopting a paraconsistent logic can also dissolve the supposed paradox by taking its conclusion to be a case of a "true contradiction".

*Sofia University "St. Kliment Ohridski"; e-mail: bmollov@phls.uni-sofia.bg; ORCID ID: https://orcid.org/0000-0002-7614-2617

i.e., resolved as instantiating argumentative fallaciousness (falsity of a premise included), is to be viewed as genuine.

For those disinclined to such a summary dismissal, there is the question of the truth-valuation or plausibility-valuation of the premises or the conclusion. In the absence of facts of the matter that can conclusively determine the truth (plausibility) value of the premises and the conclusion, as those are usually non-factual (non-empirical), any valuation could be accused of arbitrariness and questioned. The more serious issue here can be that it is precisely the non-empirical (non-empirically verifiable) character – and the resultant semantic and conceptual vagueness – of propositions used as premises or conclusion that makes paradoxes possible. To put it in a Quinean-esque slogan form, "No paradoxicality without vagueness". Eliminating vagueness would indeed resolve (dissolve) paradoxes; yet, again, that would be of rhetorical interest only, since the resolution proceeds by way of revealing the semantic fallaciousness (that vagueness is) beneath the appearances of paradoxicality, i.e., by identifying them as pseudo-paradoxes.

Actually, being resolvable at all can be taken to indicate pseudo-paradoxicality. (A)-resolvable paradoxes are disqualified as instances of argumentative or semantic fallaciousness that justifies being studied only with a view of knowing how to avoid it. And although (b)-resolvable paradoxes seem to substantively contribute to knowledge by adding to it what was arrived at through paradoxical reasoning, they obviously cannot be counted as paradoxes since their conclusions are no longer considered what by definition they should be: counter-intuitive, counter-factual, or false. If an epistemic insight of that sort is nevertheless seen as valuable, because, paradox or not, our positive knowledge gets extended anyway, there is the concern that (b)-resolvable paradoxes are in fact cases of *petitio principii* and so are also fallacious if given the form of an argument, i.e., whatever epistemic value a paradox apparently has is contained in one or more of its premises that are – by definition – sufficiently plausible (intuitive). Since the (plausible) premise presupposes the conclusion, the paradoxical argument is redundant unless we take it to be a special form of analysis that can force us to recognize the implications of certain pieces of pre-existing knowledge, but why call it a paradox then?

It may be objected that showing that and how apparent paradoxes are actually pseudo-paradoxes is what resolving paradoxes amounts to precisely. And that this is the only resolution there is or can be expected to be, which does not deprive paradoxes of a conceptual, or logical, linguistic, rhetorical, paedagogical, and/ or stylistic value any of which is still significant enough to appreciate. While I agree that these functions do carry considerable value, no matter how inflationary one's meta-philosophy is, there do exist expectations (shared by Chisholm too) that paradoxes are capable of substantive contributions to metaphysics. Again, no matter how deflationary one's meta-philosophy is, a solution to what seem to be substantive problems by way of showing that there were no problems to begin with but only fallacious reasoning and conceptual confusion, strikes as deeply and unnecessarily underwhelming. There may indeed be no substantive metaphysical lessons to be learned from paradoxes apart from how not to commit certain fallacies, as I argue here, but that should not be made a matter of meaning, of defining "paradox resolution" in a way that discourages expectations of metaphysical uses for them.

The only suitable candidate for genuine paradoxicality that we appear to be left with is Quine's third class of paradoxes, "antinomies". Or these that look like "unequivocally" falsidical paradoxes in that their conclusions resist attempts to be found (made) plausible (and so are not (b)-resolvable), but show no structural fallaciousness or premise falsity (and so are not (a)-resolvable either). Hence, according to Quine, the resolution (elimination) of "antinomies" requires a revision of our conceptual scheme, which, for him, quite deflationarily, involves piecemeal amendments to the logical or natural language in which they have come to be derived, his examples being Grelling's (heterological adjectives) paradox, Epimenides' (Liar) paradox, Berry's ("numbers and syllables") paradox, Russell's (class self-membership) paradox and Cantor's (power class) paradox. Thus, Quinean antinomies, rather than leading to some revolutionary overhaul of our conceptual scheme, turn out to be solvable by suitable linguistic revisions and so can also be disqualified as either redundant or instances of fallacies (linguistic – lexical or syntactic – vagueness)[3].

The further issue is that those non-fallacious or otherwise non-solvable paradoxes, specimens of which are still to be discovered, would be of no conceptual or metaphysical use. Since they cannot be solved in principle, there is no way to extract whatever metaphysical or conceptual lesson they may, in their "unfathomable mysteriousness", hold. Thus, we face what I was tempted to call the "paradox of paradoxes", though given its nature, "puzzle" would be more appropriate: the puzzle that paradoxes are either solvable (as fallacious) and hence pseudo-paradoxes, or are not solvable and hence uninformative.

3. Modal paradoxes

Modal paradoxes put the mentioned issues into especially sharp relief. Take the type of modal tolerance paradoxes usually referred to as "Chisholm's modal paradox" (CMP) (Chisholm, 1967, 1973). These commonly exhibit a straightforward argumentative structure, validly deriving a conclusion of modal hyper-tolerance (MHT) from the premises of modal tolerance (MT) ($[(x)(P_N x \rightarrow \Diamond P_{N+/-1}x)]$) and iteration (I) ($[\Diamond\Diamond\phi \rightarrow \Diamond\phi]$) with the additional premise of (MT)'s necessity (NMT) ($[\Box(x)(P_N x \rightarrow \Diamond P_{N+/-1}x)]$)[4]. I.e., the premise that an object could have been slightly different in a given respect (MT) and the premise

[3]It has also been suggested that the Quinean taxonomy is problematic mainly due to its "heavy dependence on the current state of one's knowledge and on one's ability to figure things out" (Lycan, 2010, pp. 617–618). The objection seems to be misguided though, because what other basis can our reasoning have? Apparently not a past/ future/ possible state of knowledge. We wish! What Lycan has in mind is that intuitions, opinions, and beliefs, including scientific ones, vary (in time or from person/ community to person/ community), so choosing which of the sets of the propositions comprising a given paradox to accept or deny will also vary, and so will the identification of paradoxes as "falsidical" or "veridical". So, Quine's classification may not always guarantee agreement as to how to classify a given piece of apparently paradoxical reasoning, but it is not clear how that can be seen as problematic as that is rather common for philosophical taxonomies.

[4](NMT) is needed to guarantee (I), i.e., the validity of (MT) at *every* relevant possible world. The paradoxicality of CP requires that (MT) is true at $W_@$, and at world W_1 possible relative to (accessible from) $W_@$, and also at W_2 possible relative to (accessible from) W_1, so there is a possibly possible world W_2 at which the given $W_@$-object (Tokyo Skytree) is 2m shorter. And there is a world W_3 possible relative to (accessible from) W_2, at which Tokyo Skytree is 3m shorter, and a world W_4 ... and so on, until we reach a world, W_n (W_{633}) possible relative to W_{n-1} (W_{632}), where Tokyo Skytree is 1m tall. If we take the relation of accessibility (of being possible relative to) to be transitive (what (I) claims), (W_{633}) would be possible relative to $W_@$, and (MHT) is entailed.

that what is possibly possible is possible (I) seem to entail the conclusion that the object could have been arbitrarily different in the given respect (MHT) [$\Diamond P_M x$]. For example, from the intuitive possibility for Tokyo Skytree Tower to have been a meter shorter (MT), and a chain of similar intermediate accessible (633) possible worlds of slight height changes (I), we are led to the intuitively absurd conclusion that Tokyo Skytree could have been 1m tall (MHT).

Chisholm's initial thought experiment (Chisholm, 1967) presents a scenario in which, assuming trans-world identity, two actual world individuals, Adam and Noah – via gradual incremental changes over a series of possible worlds $W^1 \ldots W^N$ – end up completely swapping their qualitative profiles[5]. The incremental changes lead to a possible world W^N that is indistinguishable from the actual world $W@$ except for the swap (Adam having all the qualitative properties Noah has in $W@$, and Noah having all the qualitative properties Adam has in $W@$), which, for Chisholm, raises supposedly "paradox-symptomatic" questions:

Should we say of the Adam of W^n that he is identical with the Noah of W^1 and should we say of the Noah of W^n that he is identical with the Adam of W^1? In other words, is there an x such that x is Adam in W^1 and x is Noah in W^n, and is there a y such that y is Noah in W^1 and y is Adam in W^n? And how are we to decide? ... [A]ren't the two Adams, the two Noahs, and the two worlds indiscernible? (Chisholm, 1967, pp. 3–4).

Chisholm 1973 sketches an analogous scenario in which two objects (tables) x and y gradually exchange parts (rather than properties) over a series of possible worlds $W^1 \ldots W^N$, ending up with a world W^M at which x is composed of all the $W@$ parts of y and y is composed of all the $W@$ parts of x, W^M being distinguishable from $W@$ solely by this parts swap. So, paradoxically, it does seem that one "can change anything to anything by easy stages through a connecting series of possible worlds" (Quine, 1976, p. 861), a highly counter-intuitive conclusion1 from the highly intuitive premise that objects' identity can tolerate slight gradual changes. The question of why we should care about intuitiveness

[5]Starting with a possible world W^2, in which Adam (930) lives a year longer than he does in the actual world $W@$ and Noah (950) lives a year less than he does in the actual world $W@$, Chisholm describes the route to W^N as follows:

> moving from one possible world to another, but keeping our fingers, so to speak, on the same two entities, we arrive at a world in which Noah lives for 930 years and Adam for 950 ... Now let us continue on to still other possible worlds and allow them to exchange still other properties. We will imagine a possible world in which they have exchanged the first letters of their names, then one in which they have exchanged the second, then one in which they have exchanged the fourth, with the result that Adam in this new possible world will be called 'Noah' and Noah 'Adam'. Proceeding in this way, we arrive finally at a possible world W^n which would seem to be exactly like our present world W^1, except for the fact that the Adam of Wn may be traced back to the Noah of W^1 and the Noah of W^n may be traced back to the Adam of W^1. (Chisholm 1967, p. 3)

What makes that paradoxical modal trajectory possible is the commitment to the belief that if Adam and Noah could have been slightly different from what they actually are, they could tolerate further slight changes to the individuals they could have been at any of the relevant series of possible worlds (i.e., that Adam and Noah could have been such that they could have been such that they could have been such that and so on to a world W^N differing from $W@$ solely with regard to which individual has which qualitative profile). Once we concede that (concede (I)), the conclusion (MHT) is inevitable and it will entail haecceitism.

carries less response value here than it does in non-modal contexts as modal metaphysics is mostly driven by intuitions; thus, CMP^2 appears to represent a genuine paradox.

4. Indicative CP solutions

Since the argument is valid, resolving the paradox would require either the acceptance of (MHT) (making it an instance of a veridical paradox) or the rejection of (MT) or (I) (making it a falsidical one). Rejecting (MT) is arguably the most counter-intuitive option, followed by the acceptance of (MHT) and the rejection of (I). Hence, the usual response to the paradox has been the rejection of (I), the so-called "intransitive accessibility" solution[6].

An alternative, "degree-theoretic solution"[7] (often conflated with the "counterpart-theoretic solution"[8]), proceeds by treating CMP as a sorites paradox, and (a)-solves it by showing that it is "degree-theoretically" invalid. The solution is unconventional as it proposes that the vague predicates generating sorites paradoxes should be taken to be satisfiable in degrees, with respective degrees of truth to be assigned to sentences containing vague predicates[9]. Since "degree-theoretical" validity is defined as a matter of making sure the degree of truth of the conclusion does not "fall below the [truth value] of the lowest-valued premises" (Forbes, 1985, p. 171), the resultant "degree-theoretical" invalidity of the modus ponens used in sorites arguments would invalidate those arguments[10]. Thus, this "many-valued" approach views CMP as "falsidical", an (a)-dissolvable pseudo-argument.

[6]Suggested by Salmon, 1986.

[7]Cf. Forbes, 1983, 1984, 1985.

[8]Cf. Lewis, 1986.

[9]In our example, "being tall" is satisfied to a higher degree by Tokyo Skytree than it is by Canton Tower or by CN Tower. Respectively, that would make the statement "Tokyo Skytree is tall" true to a higher degree than "Canton Tower is tall" and both true to a higher degree than "CN Tower is tall." Statements containing vague predicates would have a range of truth values beside "completely" true or "completely" false (though the possibility of the "completeness" of these two values for vague predicates can be questioned). It is the failure to recognize the truth value range that – according to the "degree-theoretic" account – gives rise to sorites paradoxes. Similarly, for the modal CP case, "could have been 10m shorter" is satisfied to varying degrees by Tokyo Skytree in different possible worlds (we will come back to that point). Consequently, "Tokyo Skytree could have been 10m shorter" at $w@$ (where Tokyo Skytree is 634m) is true to a higher degree than "Tokyo Skytree could have been 10m shorter" at w_1 (where Tokyo Skytree is 624m), true to a higher degree than "Tokyo Skytree could have been 10m shorter" at w_2 (where Tokyo Skytree is 614m) and so on, which would make CMP invalid since the conclusion would be true to a lower degree than the lowest-valued of the premises.

[10]Modus ponens being the only inference pattern employed in sorites arguments. Forbes's "many-valued" truth-table clause for the conditional is as follows:

$Val[A \rightarrow B] = 1 - (Val[A] - Val[B])$, if $Val[A] > Val[B]$ (Forbes 1985, p. 171).

Thus, modus ponens would not be a valid pattern of inference, for example, say, the antecedent in the conditional premise has the value 0.5 and the consequent 0.3, then the conclusion's value (0.3) would be lower than that of the premises (respectively, 0.8 for the conditional premise, and 0.5 for the antecedent premise), making it invalid.

And so do the (I)-denial solutions[11], for which CMP is unsound due to the supposed falsity of the (I)-premise. These solutions claim to properly reflect the modal intuition that it is not the case that what is possibly possible is possible, i.e., that the accessibility relation between possible worlds is not transitive: if α that is F in a possible world W_1 could (in W_1) be G, and if α that is G (in W_2) could (in W_2) be H, it need not be the case that α in W_1 could be H.[12] The rejection of (I), as a false premise, would block (MHT) and (a)-dissolve CMP, but that raises the question of whether CMP should have been classified as a paradox at all. Given that paradoxicality is taken to be produced by the tension between highly plausible premises and unacceptable conclusion, CMP's controversial (I)-premise seems to disqualify the scenario from the status of "paradox". The intuitions about iterated modalities are usually too weak to meet the "high plausibility" requirement and in the case of (I) they even go the other way, against it. And it is not that the lack of sufficient plausibility can be made up for by the availability of an argument that can provide (sufficient) support for (I), for no such argument exists.

Turning to the more counter-intuitive options, though the conviction that any object could vary in its – at least, non-essential – qualitative characteristics seems to be overwhelmingly obvious, denying (MT) may not be that absurd an option for (a)-dissolving ("falsidicalizing") CMP. Admittedly, the hyper-essentialism it is based on, is similarly counter-intuitive (or is it?), but is not indefensible. It is a key component of the counterpart-theoretic solution as well as of Forbes's "degree-theoretic" solution[13], with both being motivated by the incoherence of the alternative of trans-world identity. The point here is that denying the (MT) premise would make CP a "falsidical" (pseudo-) paradox.

A graver problem with (MT) is that it makes CMP argumentatively unsound, a *petitio* rather than the *reductio* (of haecceitism[14]) it was supposed to be. (MT) entails anti-essentialism, since if a property Pe that α does not have at some world is necessary for its identity (is one of the set of α's essential properties), as essentialism would hold, then (MT) would be invalid. Assuming (MT), at some worlds it would not be possible for α to

[11]The (I)-denial comes in two forms: Salmon's intransitivity of the accessibility relation and Lewis's intransitivity of the counterpart relation. Their difference does not affect the conclusion drawn here. Salmon thinks that there is a fallacy in (I): "Thus the modal paradoxes turn on a fallacy special to S4 modal logic. In deriving the paradoxes in S4, one commits the fallacy of possibility deletion, inferring $[\Diamond\phi]$ from $[\Diamond\Diamond\phi]$ or equivalently, the fallacy of necessity iteration, inferring $[\Box\Box\phi]$ from $[\Box\phi]$" (Salmon, 1986, p. 82).

[12]Or the possible world W_3 that is accessible from W_2 that is accessible from W_1 need not be accessible from W_1.

[13]Since for Forbes, in the modal case, the "predicate whose application conditions are tolerant is a predicate expressing the thisness or haecceity of α, $\xi = \alpha$" (Forbes, 1983, p. 239). But speaking of degrees of satisfaction of $\xi = \alpha$ would require accepting degrees of identity for α, which Forbes thinks is an "incoherent notion". Therefore, he proposes that the trans-world identity relation of standard S5 semantics is replaced with the (trans-world) similarity relation of the counterpart theory as the idea of degrees of similarity is unproblematic (ibid., pp. 247 – 248).

[14]Haecceitism is usually considered a form of anti-essentialism in that it denies that there are essential *qualitative* properties that objects must have in order to be what they are (e. g., Nefertiti could have been anyone and anything: Nina Simone or Newton, a nightingale, a nectarine, a needle). If, however, the properties that an object needs to have to keep its identity are not restricted to only qualitative ones, haecceitism would be a version of essentialism, since it would require an object to essentialistically have a haecceity ("being itself") or other haecceitistic (identity) non-qualitative properties (e. g., "being distinct from X", "being the same age as X") at all possible worlds at which that object exists.

be Pe, because of the gradual changes $P_{N-/+1}$ in the property profile of α that (MT) allows, which would imply that if Pe is considered to be part of α's essence, then α would not exist at the worlds in which α is not Pe (as per essentialism) and therefore (MT) cannot be true. I. e., if essentialism, then non-(MT); and if (MT), then non-essentialism. But if (MT) and (NMT) are incompatible with essentialism, CMP would be fallacious ((a)-dissolvable) since it directly begs the question against essentialism (by presupposing the anti-essentialist conclusion it claims to have argued for).

Accepting (the highly counterintuitive[15]) (MHT) would make CMP "veridical", (b)-solvable, but if (MHT) can be accepted with or without additional consideration, even when the latter has been provoked by the appearance of the paradox, then its paradox-generating character of being inherently, overwhelmingly, etc. unacceptable would be lost and with it the reason to identify paradoxicality in the scenario.

CMP is redundant since accepting (MT) would inevitably commit to (HMT), respectively, rejecting (MT) or (I) eliminates the paradox: no argument is needed to establish or negate (HMT). Indicatively, it is not even clear what the argument is. Assuming that it is meant to be a *reductio*, what is it a *reductio* of: haecceitism or anti-haecceitism? It can easily go either way, depending on one's modal metaphysical views or intuitions about its premises and conclusion (and accepting (HMT) can also be a "solution"). Any of the options – rejecting or accepting (MT), (I), (MHT) – would suffice for the respective ontological claim; there is nothing in the paradox that goes beyond or adds to the content of the premises, so presenting CMP as an argument for a claim is misleading and would make it fallacious (a *petitio*) anyway. In that sense, CMP does more metaphysical harm than the arguable rhetorical good it may be taken to bring (say, highlighting or explicating the metaphysical commitments of the premises). Ironically, what is most detrimental in presenting it as a paradoxical argument is precisely the way that particular rhetorical form or device – in imaging contrived hypothetical scenarios – can hide or distract from those very metaphysical commitments it is supposed to elucidate. Admittedly, that does not apply to all thought experiments, and, hopefully, that restricted rhetorical use of paradoxes can be further regulated to be made more beneficial.

5. Conclusion

Since modal "paradoxes" seem to depend entirely on the intuitions underlying their premises and conclusions, any variation in those would affect their paradox status. As CMP shows, one modal metaphysician's paradox is another's quotidian conceptual commute. If the high hope is for paradoxes to bring about radical, at least partial, revisions of current conceptual schemes, it may be a faint, if not a false, one. As mentioned, the puzzle is particularly pronounced in the modal case: paradoxes are either generated by fallacies or are redundant, or if neither, given we manage to find ones of this elusive third kind, would be unenlightening.

[15]The dominant modal intuition is non-(MHT), or that the modal profiles of objects do not permit radical (complete) variations. One popular restriction on the extent of objects' modal variability is Kripke's origin essentialism (Kripke, 1980), the plausible but not universal view that the origin of an object is essential to it, i. e., the object's identity does not tolerate a radically or completely different origin from the origin the object in fact has. Another popular restriction is sortal essentialism, the view that an object could not have been of a radically or completely different kind from the kind that the object in fact is.

The question as to why cases of the third ("antinomic") kind have been so hard to encounter is probably worth pursuing elsewhere, better in comparison with similar elusive types of entities, but one answer would be that we have not been able to find them because they do not exist. I.e., paradoxes are generated from within the conceptual scheme and the linguistic framework we have and use, usually as a result of a fallacious use of those resources or a failure to properly understand their capacities. This indicates a possible (deflationary) role for paradoxes: they may be expected to reveal, analyze, account for, and correct cases of misuse of conceptual and linguistic resources. Probably overoptimistic – for the same reason "antinomic" paradoxes evade us: meta-perspectives are not entirely meta, as they are still accessed within the limits of the object conceptual/linguistic framework – but an option anyway.

As mentioned, CMP can be taken to present a case for the argumentative use of paradoxes since the thought experiment was designed and has been since interpreted as a *reductio ad absurdum* of haecceitism. But then again, the crucial (I)-premise is too weak to bear the burden of the argument, as arguably are (MT) and not-MHT. Intuitions, especially modal ones, do vary significantly and cannot be tasked with an argumentative responsibility of that scale, which does not mean that in certain instances they cannot be put to rhetorical use. That may not be the substantive philosophical revisionist use expected of antinomies and it is unlikely that pure rhetorical devices can sway anyone's metaphysical views, but paradoxes can be instrumental in making certain modal points more clearly, (hopefully) more precisely, or at least, and what no reader of current philosophical prose can possibly object to, stylistically more appealing.

Acknowledgments:

This study is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT BG-RRP-2.004-0008-C01. The author would like to thank Alexander Kanev for numerous discussions on the topics of the paper.

References

Chisholm, R. M. (1967). Identity through possible worlds: Some questions. *Noûs*, 1(1), 1–8.

Chisholm, R. M. (1973). Parts as essential to their wholes. *Review of Metaphysics*, 26(4), 581–603.

Forbes, G. (1983). Thisness and Vagueness. *Synthese*, 54(2), 235–259.

Forbes, G. (1984). Two Solutions to Chisholm's Paradox. *Philosophical Studies*, 46(2), 171–187.

Forbes, G. (1985). *The Metaphysics of Modality*. Oxford: Clarendon Press.

Kripke, S. (1980). *Naming and necessity*. Cambridge: Harvard University Press.

Lewis, D. (1986). *On the Plurality of Worlds*. Oxford: Blackwell.

Lycan, W. (2010). What, exactly, is a paradox? *Analysis*, 70(4), 615–622.

Quine, W. V. (1966). The ways of paradox. In *The Ways of Paradox and Other Essays* (pp. 3–20). New York, Random House.

Quine, W. V. (1976). Worlds away. *The Journal of Philosophy*, 73(22), 859–863.

Salmon, N. (1986). Modal paradox: Parts and counterparts, points and counterpoints. *Midwest Studies in Philosophy*, 11(1), 75–120.

Balkan Journal of Philosophy (2025), vol. 17, issue 2, 146–164

HOW CAUSAL REALISM SURVIVED RUSSELL'S ATTACK: OPPORTUNITIES TAKEN AND LESSONS LEARNED

Milan Z. Jovanović*

Đorđe Lazarević**

Abstract: In his seminal paper *On the Notion of the Cause*, Bertrand Russell famously attempts to show why the view that causation is a real and mind-independent feature of the world – called "causal realism" – is false. His arguments are often cited as a definitive challenge to the metaphysical tradition that takes causality as a fundamental principle of reality. Yet, despite the apparent plausibility and power of his arguments, causal realism is still a predominant view within philosophy of causation today. This paper seeks to explore how causal realism has endured Russell's attack, examining not only the weaknesses in his arguments, but also the ways in which contemporary philosophers have redefined and strengthened the realist position. By analyzing the responses to Russell's arguments, we will try to pinpoint the lessons learned from his critique and demonstrate how it has informed the development of more sophisticated theories of causation. Doing so will also contribute to a redefinition of causal realism, showing how the debate with Russell has enriched our understanding of what it means for causal theory to be realistic.

Keywords: Bertrand Russell, causation in science, causal realism, Lewis' counterfactual theory of causation, temporal asymmetry.

1. Introduction

Bertrand Russell's *On the Notion of Cause* stands as one of the most influential and radical critiques of the traditional philosophical understanding of causality. Written from a naturalistic perspective that acknowledges the authority of science in metaphysical inquiry, Russell's paper is typically understood as an attack on causal realism – the view that causation is an objective feature of the world, independent of human language and cognition. His arguments are often cited as a definitive challenge to the metaphysical tradition that takes causality as a fundamental principle of reality.

Yet, despite the apparent plausibility and power of his arguments, causal realism has not only survived but has also evolved in response to his critique. It still is a prominent view within philosophy of causation today.[1] This paper seeks to explore how causal realism

[1]It should be noted, however, that the study of causality has undergone a significant transformation – or even a *revolution* (Pearl & Mackenzie, 2019) – in recent decades, with developments that appear favorable to causal antirealism. Among the most influential recent theories are those that analyze causation using statistical models, graphs, and structural equations (e.g., Halpern & Pearl, 2005, Pearl, 2000, Spirtes, Glymour, & Scheines, 1993), which are at least difficult to reconcile with causal realism in the relevant sense. That said, these approaches do not so much endorse antirealism as they shift focus away from metaphysical

*Faculty of Philosophy, University of Niš, Serbia; milan.jovanovic@filfak.ni.ac.rs

**Faculty of Philosophy, University of Niš, Serbia; djordje.lazarevic@filfak.ni.ac.rs

has endured Russell's attack, examining not only the weaknesses in his arguments, but also the ways in which contemporary philosophers have redefined and strengthened the realist position.

By analyzing the responses to Russell's arguments, we will try to pinpoint the lessons learned from his critique and demonstrate how it has informed the development of more sophisticated theories of causation. Doing so will also contribute to a redefinition of causal realism, showing how the debate with Russell has enriched our understanding of what it means for causal theory to be *realistic*.

2. Russell's arguments against causal realism

The arguments offered in favor of Russell's radical conclusion are very diverse, and they vary with respect to the aspect of causality they are directed at, the strategy they are employing, the nature of the premises they rely on, and so on. However, we are not going to analyze Russell's argumentation in its entirety. A substantial part of Russell's objections is specifically directed to the questions of will, agency, determinism, and teleological systems – the questions, although relevant to his enquiry, still somehow peripheral to our investigation. Instead, since we are primarily concerned with causal realism, in what follows, we will be focused on the arguments typically recognized as central to his critique of the law of causation and causal realism.

As it is commonly assumed,[2] there can be identified two such arguments in Russell's paper that are directly relevant to causal realism (we will call them: A_1 and A_2, respectively). The first one (A_1) is implicit and consists of two separate objections – the first of which (O_1) is at the very beginning of Russell's paper, while the other (O_2) is somewhere in the middle.[3] The second argument (A_2) is more explicitly formulated and thoroughly discussed towards the end of the first third of the paper (Russell, 1959, pp. 185–188).

2.1. Russell's first argument

Let us start the analysis by considering the first objection within the first argument. O_1 is directed against (then) widespread philosophical conviction that – since every event has a cause – the main goal of natural science (as an enquiry into objective reality and its structure) is to look for causes. To that claim, which somehow transitioned from modern metaphysics to contemporary philosophy of science, and which puts an *a priori* restriction

debates, emphasizing instead operational definitions, causal modeling, inference, methodological concerns and applications.

[2]Most interpreters recognize that there are at least two anti-realist arguments in Russell's paper, but they differ with respect to specific formulations they offer, emphasis and relative strength of the arguments. For example, N. K. Shinod (2019) and Alyssa Ney (2009) take both A_1 and A_2 primarily as arguments against causal realism in *science*, with Shinod arguing for success of both arguments, and Ney for A_1 as being more persuasive, but still inconclusive. On the other hand, Corry (2006), Hitchcock (2007), Menzies (2007) and Eagle (2007) consider both, or at least A_2, as having a broader target of causal realism *per se*. In the interpretation that follows, we agree with the latter group with regards to the scope of the arguments and maintain that it is the second argument, A_2, that is stronger.

[3]In interpreting or rather "parsing" Russell's argumentation in this way, we are following Jovanović (2020, pp. 3–6).

on empirical science,[4] Russell responds by stating that "oddly enough, in advanced sciences such as gravitational astronomy, the word *cause* never occurs" (1959, p. 180).

Although it is easy to see how O_1 objects to the commonplace philosophical conception of science, it is nevertheless questionable how *plausible* it is *against*, or even *relevant to* the philosophical usage of the word cause (outside of the philosophy of science). Even if Russell is right in proposing that there is no causation in fundamental physics,[5] it is not apparent how that would disqualify the notion of cause as sensible and definable (which seems to be Russell's ultimate goal). Though it may be true that causation is not, to the best of our scientific knowledge, the part of the world inventory, that still does not prove impossible a project aiming to define causation (as a distinct relation) *on the* relations and entities recognized as the constituents of the fundamental reality by the best scientific theories. In that way, not only could the notion of the cause turn out to be meaningful and precisely defined, but it could also be claimed to be *real*, i.e. to be adequate for the position of *causal realism* – in the sense yet to be elaborated.[6]

But this is only a partial presentation of Russell's first argument. In the second objection,[7] he strengthens the critical comment just considered with philosophically more relevant insights. In O_2 Russell highlights that the theories of advanced fundamental sciences are not formulated in terms of causes and effects, but instead they state their laws by the means of specific equations. He writes:

> In the motions of mutually gravitating bodies, there is nothing that can be called a cause, and nothing that can be called an effect; there is merely a formula. Certain differential equations can be found, which hold at every instant for every particle of the system, and which, given the configuration and velocities at one instant, or the configurations at two instants, render the configuration at any other earlier or later instant theoretically calculable. That is to say, the configuration at any instant is a function of that instant and the configurations at two given instants. This statement holds throughout physics, and not only in the special case of gravitation. But there is nothing

[4] In his paper "Causation as Folk Science", John Norton offers a detailed argument against causal realism based on the problematic *apriority* of the causality principle – as typically formulated in philosophy, See Norton (2007).

[5] If the interpretation that Russell targets only causation in science is correct, as suggested by authors such as Loewer (2007) and Shinod (2019), Russell's response to this observation of the potential irrelevance of O_1 for causal realism would be that *this* is irrelevant to *his* argument. Alyssa Ney ascribes a slightly different view to Russell, according to which he follows the thesis she calls "foundationalism about causation." This is the claim that the existence of causes on any level of reality presupposes their existence on the fundamental level; and, since Russell argues against the latter, we should disregard the notion of causation in its entirety (Ney, 2009, p. 8). In what follows, we pursue a different interpretive strategy from that of Loewer and Shinod, and argue against Ney's foundationalism about causation thesis (see Section 2).

[6] This is an important point. Why this should be considered a kind of causal realism will be discussed in the second and third part of the paper.

[7] Corry and Ney regard the passage (that we labeled second objection) as the separate (and complete) argument and call it "the asymmetry argument" (Corry 2006, p. 263), or "the directionality argument" (Ney 2009, p. 11). Similarly, within our O_2 Blanchard – roughly following Field (2003) – recognizes two separate arguments: *the globality* and *the symmetry* argument (2016, p. 257).

> that could be properly called cause and nothing that could be properly called *effect* in such a system (1959, p. 194).

This critical remark stresses a worrisome structural "mismatch" between (even the broadest and most abstract) conception of causation in philosophy and the physical framework to which it should (according to the naturalistic approach endorsed by Russell) somehow coincide. It is typically assumed in contemporary philosophy of causation that causation is an asymmetric relation – closely related to determination, and closely tied to the natural laws – between a cause and an effect, often understood as events in temporal order in accordance with the direction of time (arrow of time). Events, themselves, are typically portrayed as some broadly construed happenings with certain duration, spatial occupation, and so on.

On the other hand, dynamical systems (similar to those explored by J. C. Maxwell) – as the best scientific description of what the philosophical notion of cause *potentially* refers to – evidently *do not coincide* with the aforementioned conception of causation (in a few relevant and significant respects). The equations expressing the laws in such systems do not support (or assume) the highlighted asymmetry inherent to the philosophical idea of causation. Instead, differential equations describing a dynamical system are *bideterministic* and perfectly indifferent with regards to the temporal dimension of its variables. Even worse, in contrast to the events as causal relata, those equations – as Russell points out later in his text – characterize *states of the system* at a given moment, by precisely specifying different physical properties of elementary particles (within the system). So, to the best of our scientific knowledge "causal relata" are precisely characterized (not loosely construed) instances (not intervals) of a given system. Hence, even the "best candidate" for causation within the fundamental science is far from being in accordance with the minimal philosophical conception of that relation.

This second objection complements the first one; O_2 picks up where O_1 left off – so to say. If a traditional metaphysician of causation responds to O_1 by stressing (as suggested earlier) that causation *doesn't have to be* a part of fundamental science in order to be *real*, then she commits herself to the exact position which O_2 is plausibly arguing against. This is the reason why these two objections should be seen to constitute *one* argument, which we will call A_1. That argument (A_1), compiled from O_1 and O_2, could be formulated like this:

Advanced science is our best guide to what is real. In theories of advanced science there is no such thing as causation; so, causation in philosophy does not simply refer to causation in science. Regularity and lawfulness of the phenomena within the system (which might be understood to give rise to a philosophical conception of causation) are best described by the natural laws which are bideterministic, symmetrical in the relevant sense, and have instances of the system's configurations as its relata. Given that *whatever* the philosophical notion of causation *refers to* seems to be an asymmetrical relation between broadly described events (seen as intervals), it is evident that *it is not referring* to those natural laws and regularity. Thus, within the philosophical vocabulary, the word "cause" does not refer (directly or indirectly) to anything real.

2.2. Russell's second argument

Nonetheless, Russell's key objection to causal realism is presented by his second argument (A_2).[8] His initial task here is to determine *what exactly* the principle of causality asserts or posits. After analyzing definitions of causal terms (using Baldwin's dictionary), Russell proceeds by supplementing those insights by citing relevant philosophical literature. He considers Mill's and Bergson's definitions of causation and notes that despite some (more or less subtle) idiosyncrasies, there is an evident consensus among philosophers regarding the principle of causality, which to some extent aligns with his analysis of dictionary definitions. However, Russell does not offer a specific formulation of that principle. Instead, he directs his forthcoming critique towards the implied generalization that a reader is expected to make from the various cited formulations (and his interventions on them).[9] For the purposes of this text, the implied version of the principle of causality could be formulated in a relatively uncontroversial manner as follows:

> (R_{cp}) Every event e is determined by some preceding event c, in the sense that all events resembling this earlier event (c), while subject to the laws of nature, are always (or necessarily) accompanied by events resembling that later event (e).[10]

The central components of the principle of causality are events, and the relation of determination (or rather: regularity). And it is exactly what Russell *aims at* with his second argument. It is evident from R_{cp} that an event is regarded as something that *can be repeated*. This implies that we would not consider an event to be *the state of the universe at a given time*. Such an event would either be unique, or the possibility of its recurrence would be negligible. Therefore, an event must be defined less specifically to avoid overly restrictive conditions for its (re)occurrence. Indeed, this is why we usually talk about events in such a way that they include things like striking a match or inserting a coin into a vending machine, without specifying the exact force with which the match was struck, or the force with which the coin was inserted into the machine, or the temperature of the coin, as well as without accounting for various circumstances (location, time, the person who did it, the hand used, etc.).

[8] This argument is typically presented as a dilemma, as we do here, although it is named differently depending on which facet of the argument is most relevant to the broader discussion. For example, Ney calls it "the localization argument" (2009) and Corry, perhaps most informatively, "the intervention argument" (2006).

[9] Hitchcock (2005, p. 46) argues that the reader is left with the impression that there are at least *three* targets of Russell's critiques: the notion of the cause, the word "cause", and the existence of causes. This is problematic because it isn't clear which argument is supposed to refute which of these targets. However, as Shinod notes, a charitable and most promising interpretation is to take Russell as arguing against the existence of causes, and that he assumes the principle R_{cp} – or something close to it – as the best approximation of the principle of causality normally presupposed by philosophical traditions. See Shinod (2019, p. 87).

[10] In this formulation, we have deliberately avoided engaging with various complexities – particularly the issue of a time interval existence between cause and effect – that Russell examines in his analysis of different formulations of the principle of causality. Additionally, following various philosophical articulations of the principle, as well as the maxim "same causes, same effects", Russell formulates the principle in terms of event types rather than token events. However, this simplification should have no bearing on the broader argument of this paper.

However, such a broad definition of causal relata brings us to another problem. If we define events in this way, if we characterize some given cause in this manner, then we *have not done enough to ensure* the occurrence of the effect. Although we take it that striking a match causes it to ignite, it is still true that not every strike of a match will be accompanied by ignition, as it can happen that the match is wet, or there is no oxygen in the room, or the angle at which it is struck over the rough surface is inadequate, or the surface itself is not rough enough to produce friction, and so on. There is no easy way out of this problem, which Russell describes as follows:

> But this means that the supposed cause is not, by itself, adequate to insure the effect. And as soon as we include the environment, the probability of repetition is diminished, until at last, when the whole environment is included, the probability of repetition becomes almost nil. (1959, p. 187)

So, we are faced with a dilemma in which both alternatives are unacceptable (if we are to preserve R_{cp}). If we take a cause to be a typically defined event – which means, as mentioned above, broadly specified and without including the circumstances – then it ceases to be sufficient for the occurrence of the effect, i.e., the regularity that underlies the principle of causality disappears. If, on the other hand, we wish to preserve that regularity, then we must take causes as highly specific events that include all the circumstances (accounting for the absence of potentially interfering factors), which leads to the conclusion that the regularity invoked by the principle of causality becomes vacuous, since there will be no instances to which it could be applied.

These two arguments form the backbone of Russell's famous attack on the principle of causality, with the clear intention of also undermining causal realism. Let us now turn briefly to the assumed *minimal* notion of causal realism, and then back to the problems with Russell's arguments and the potential response of contemporary causal realism to them.

3. Causal Realism and Modest Causal Realism

In the preceding discussion, we have examined Russell's arguments against *causal realism* with an implicit understanding of the concept. However, it is important to make this understanding explicit in order to ensure conceptual clarity. In both this work and broader philosophical discourse, causal realism is generally understood as the position that causation is an objective relation in the world, independent of human cognition or conceptual schemes. In its stronger formulation, causal realism asserts that causal relations are intrinsic to the world's ontological inventory, constituting fundamental and irreducible features of reality. However, as many of Russell's argumentative remarks suggest, causal realism can also take a *weaker form* while remaining committed to the core thesis that causation is independent, objective, and *tied to* the world.

This weaker version, which we may term *indirect causal realism* (or *light*, or *modest* causal realism), conceptualizes causation as a relation that is *defined on* fundamental entities, properties, and relations; it is not by itself fundamental, yet it is ultimately *reducible to* or *grounded in* something that is. Thus, a causal realist need not claim that causation exists as a *sui generis* entity – an approach that might be associated with certain classical

metaphysical traditions from antiquity, medieval philosophy, or early modern thought. Instead, this modest causal realism can accommodate more deflationary ontologies that explain causation in terms of facts, properties, forces, processes, etc. that are still tied to the world in some more or less uncontroversial way. Or, epistemically speaking, the modest causal realist can ground her causal theory in our best fundamental scientific theories, although those theories are themselves – *causationless*.

This more liberal or, in a sense, minimal conception of causal realism – that acknowledges both these forms of causal realism as legitimate – aligns with the working definition assumed in the previous discussion and will also serve as the framework against which Russell's objections will be evaluated in the next chapter.

Admittedly, there exists a long-standing practice of considering primarily these stronger and more committed positions as realism,[11] according to which many of the accounts we consider realist are consequently being perceived as closer to antirealism. Nevertheless, there are two important points to keep in mind here (both of which have been hinted at above). First, our approach critically depends on the notion of realism that is: a) in accordance with the proclaimed aim of Russell's critique, and simultaneously b) has a clear and coherent criterion for its application.

Regarding the first point, it is evident from his text that Russell conceives realism more broadly, offering arguments that concern this more minimal notion of causality – and, more importantly, believing that his arguments work even against these indirect forms of causal realism. As for the very notion of causal realism and the criterion for its application, it has been elaborated in the preceding paragraphs and will be used with that meaning – that is, with that specific intension – throughout the following chapter.

4. The Response of Contemporary Causal Realism

Now, with these clarifications in mind, we can ask – what is the outcome of Russell's critique? Does it succeed in achieving its stated goal? What is the epilogue regarding causal realism – is it really defeated? And what happened to the philosophical vocabulary – was causation eliminated from it?

4.1. Problems with Russell's first argument

Given that Russell appeals for changes in philosophical practice, it is useful to begin by assessing the current state of the relevant philosophical literature. The fact of the matter is that the concept of cause continues to play an equally significant role in philosophy today. Even worse (from the Russellian perspective), in contemporary philosophy of causation, realism remains the dominant position.

There are several reasons why the situation regarding causation and causal realism in contemporary philosophy is not surprising. In what follows, we will focus on those reasons that are *inherent* to Russell's critique and the problems in his argumentation that help casual realism survive.

[11] We are grateful to the anonymous reviewer for highlighting the need to clarify this point, as well as for other significant critiques addressed elsewhere in the paper.

Let us proceed step by step, starting with the first critical remark in Russell's text (O_1) – that causation is not mentioned in modern physical theories. According to some authors this claim is not true.[12] However, even if it is true that causation is not mentioned in modern physics, or were to be true, it would have only shown that: a) the philosophical speculative claim that science is essentially dedicated to investigating causation is mistaken, b) causation is not by itself (recognized as) a fundamental relation from the scientific point of view. Undoubtedly, Russell's argumentative aspirations were stronger than either of these two highlighted outcomes. He was after realism in the philosophy of causation,[13] and this is far from undermining causal realism.

To clarify, the absence of causation from fundamental natural science does not *necessarily* spell bad news for the causal realist. Even if modern physics does not use causal terms or explicitly discuss causation, it is still possible to ground (or define) causation in a way that satisfies all the desires of a causal realist – that it is objective, directly tied to the world, and independent of the human mind. As Peter Menzies (2007) observes and emphasizes, this is a serious problem for Russell's argument:

> Put this simply, the criticism is not completely persuasive. For a defender of the causal concept could well argue that, even though causation is not explicitly mentioned in fundamental physics, it is implicitly present in the picture of reality given in fundamental physics, since causal relations supervene on the pattern of fundamental physical facts and physical laws. This more sophisticated doctrine is one Russell certainly never formulated because he did not have the concept of supervenience to hand. (Menzies, 2007, pp. 191–192)

This theoretical possibility, not *fully* appreciated and covered by Russell,[14] has in fact been utilized to reformulate causal realism in contemporary philosophy of causation. By rejecting the implicit premise that causal realism requires our fundamental science to explicitly discuss causation, contemporary realists have built their theories around the concept of cause as (somehow) supervening on certain fundamental entities, properties, and facts. And that is the real reason causal realism survived Russell.

A good example of this are contemporary *process theories of causation* (Dowe, 2007; 2010). The causal relation in these theories is defined as supervening on facts about causal processes, their interactions, the persistence (Ehring, 1997) or exchange (transfer) of certain physical properties (Aronson 1971; Fair 1979) – all of which are ultimately reducible to the

[12]Patrick Suppes and Christopher Hitchcock point out that gravitational astronomy, which Russell took as an example of a theory devoid of the concept of cause, is a very specific discipline and, as such, inadequate for such a generalization. Moreover, both of them, at the time of writing their respective texts, citing the latest articles in prestigious professional journals in the field of physics, demonstrate that the concept of cause does in fact appear relatively frequently and in relatively diverse areas of physics. See Hitchcock (2007, p. 55), Suppes (1970, p. 6).

[13]Hitchcock (2007) understands the role of this critical remark in a very similar way.

[14]Nonetheless, Menzies' remark is likely too harsh, as Russell's next comment makes it clear that he, at least to some extent, understands and takes into account this variation of realism. Of course, this does not mean that he has in mind the idea of supervenience with all of its technicalities, but rather points to his awareness that realism can also be built around a causal relation that is indirectly connected to the world – one that is defined on some fundamental relations. This issue will be further elaborated in discussing O_2.

laws of modern physics. Interestingly, those theories can be seen to have their predecessor in the late Russell's theory of *causal lines* (1948). However, these accounts of causation rather elegantly avoid O_1, uncontroversially standing as causal realism without commitment to causation as fundamental entity.

Although very different from metaphysical theories mentioned above, so-called dependence theories of causation can still be seen to offer another good example of dealing with first Russell's objection. Among them we have the *counterfactual theory of causation*, that is *modestly realist* in the sense outlined earlier. According to them, causal relation is defined in terms of counterfactual dependence, which is an objective relation grounded in counterfactual conditionals – whose truth, in turn, is a function[15] of objective facts about the world and natural laws (Lewis, 1973a, 1979; Stalnaker, 1968).[16] Even probabilistic theories of causation (Hitchcock, 2018) can be seen as examples of causal realism in this sense, as long as the probabilities they operate with are defined as *objective* and *independent* of the human mind (cognition and language). In a bit more controversial manner, even Mackie's regularity theory (1974) and Woodward's *manipulability theory* (2003) could be seen as valid examples of realistic causal theory that aims to avoid Russell's objection.[17]

Although Menzies, as we have seen, notes that Russell (lacking the concept of supervenience) could not have anticipated more sophisticated versions of causal realism, it is interesting that the next critical remark in Russell's text (O_2) aims at exactly that back-up realist position. More precisely, Russell's observation about the bideterministic nature of the differential equations used in physics can be understood as a supplement to the previous critique, supporting the thesis that the concept of cause (is not only absent from the modern-day physics, but also) lacks a clearly defined content. This concept cannot directly refer to *causation in physics*, since causation is not explicitly used in physics (O_1), nor can

[15]This holds without reservation for so-called standard semantics of counterfactual conditionals, which is a unifying term for Lewis's and Stalnaker's theories (Lewis, 1973a; Stalnaker, 1968). However, it should be kept in mind that there are theorists of counterfactual conditionals who doubt that such sentences have truth values (for example: Edginton, 2008, van Fraassen, 1980) or have additional concerns regarding the semantics of counterfactual conditionals and their reference to objective reality (Price, 1997; Hájek, 2014). If a theory of causation were based on such theories of counterfactuals, it could not be properly called *causal realism*.

[16]In their analyses of the causal claims, those dependence theorists sometimes draw the line between semantic and pragmatic factors relevant to our causal statements. Although these "minor pragmatic elements in the causal concept", as Menzies cites them, can be seen to cast some doubt on the realist nature of dependence theories, authors of those theories aspire to show that "... causal concept has completely objective truth-conditions which can be stated in terms of conditions holding of the mind-independent substructure." (Menzies, 2007, p. 193)

[17]Menzies provides a more exhaustive list with a precise characterization of supervenience in each of the theories (Menzies, 2007, pp. 192–193), concluding his overview with the following comment: "Though these theories differ in detail, they all subscribe to the doctrine that causal relations depend completely on a substructure of mind-independent relations", thereby emphasizing the realist aspirations of all the mentioned theories. The situation regarding Woodward is probably a bit more complex. His position is famously non-reductive, yet closely tied to conditionals through the concept of possible interventions. Moreover, in the Introduction and chapter "Interventions, Agency, and Counterfactuals" he repeatedly insists that, according to his analysis, causal claims are *objective*, *real* and independent of our epistemic apparatus.

it refer to *what we have in physics instead of causation* – namely the regularities and pattern of dependencies expressed through physical laws (O_2) – because this regularity is, in an important sense, inadequate to support causal relation.

What motivates Russell to consider the regularity in science (primarily given through dynamical systems and expressed with differential equations) as a candidate for the basis of the concept of causation is fairly uncontroversial. There is, alongside a strong intuitive connection, a long tradition of linking regularity and causation. As mentioned earlier, Russell in his analysis adopts from the philosophical tradition the definition that causation is determination: some event c causes e when it determines it, i.e., when the occurrence of c (or events like c) together with the laws is sufficient to ensure the occurrence of e (or events like e). This is precisely why the determinism found in dynamical systems presents itself as a good candidate (or rather: best available). But, as Russell further argues, the determinism found in those specific domains of physics *lacks the temporal asymmetry* that is usually taken to be a necessary property of the causal relation. And this, according to Russell, is a serious problem. It seems apparent that this objection is essentially directed at more sophisticated theories of causal realism (indirect, light or modest causal realism) in which the causal relation is not fundamental but somehow defined on fundamental entities or properties. Especially with regards to those *indirectly realistic theories* this potential mismatch – between asymmetry that we have in causation and symmetry that we have in that potential causation bearing (or grounding) relation – becomes a pressing issue.

However, the plausibility and effectiveness of this objection should not be overstated. First, it should be noted that Russell's claim about the essentially time-symmetrical nature of evolution of physical systems in fundamental physics has been challenged in at least two respects. One direct way is to point out that physics has progressed since 1913 (when Russell's paper was originally published) and that certain symmetry-breaking phenomena are now widely postulated within the Standard Model of particle physics and quantum mechanics. For instance, time-symmetry violation in weak interactions (an experimentally confirmed fact) and the intrinsic asymmetry in the evolution of the wave function pre- and post-collapse (at least a theoretical possibility) are now part of mainstream physics (Field, 2003).

Another, philosophically more significant reply, is to notice that it is not clear if Russell is right in taking the *mathematical formalism* of such theories, even if it does imply the mentioned symmetry, as a complete description of this level of reality. For example, Esfeld argues that it is precisely *causation* that is the best candidate for a qualitative property of physical structures that distinguishes them from mathematical structures that (only quantitatively and necessarily imperfectly) describe them (Esfeld, 2011, pp. 5–6). Donald Lipkind makes the same remark when discussing this argument and brings an example from Hans Reichenbach in order to explain the difference.[18]

[18]If we imagine a light beam going from point A, passing through a red filter and arriving at point B as red light, this description of the process seems far more intuitive than the one of light being originally red and becoming white after it passes the filter, even though the second description is compatible with the time-symmetric mathematical formulation of the laws of electrodynamics. This example has its limitation, in that fundamental physics may lack an analogous "filter" or "mark" that breaks the symmetry, but, Lipkind concludes, "nevertheless, the persuasiveness of Reichenbach's interpretation of the beam of experiment points

However, let us, for the sake of argument, assume that Russell is correct that laws of fundamental physics are time-symmetrical. We can still ask why should the symmetry of a given relation of determination (as typically found in the laws of physics) be an *a priori* problem, let alone an *insurmountable* one, for the project of analyzing causation in terms of that relation? There is a possibility that this analysis need not be *entirely reducible* to the determinism given by a specific relation in fundamental physics alone; it doesn't have to be a *function* of it. The analysis in question could include other determinants that help *align it* with the properties of the (philosophically defined) causal relation. As Eagle suggests, addressing this very objection of Russell:

> ... [A]lthough it might be true that the causal asymmetry is not an asymmetry of determination, causation still might be defined from a relation of determination combined with an asymmetrical relation, where the asymmetry comes from somewhere else. (Eagle, 2007, p. 158)

And again, this is not *merely* a theoretical possibility; the theoretical space that Russell overlooks in his argumentation has indeed been utilized by contemporary realism. The asymmetry of the causal relation is achieved in various ways in contemporary causal theories. A direct way to introduce asymmetry into a system described by symmetric laws is to rely on something present in physics itself – the famous second law of thermodynamics and the principle of entropy, which is mentioned in its formulation (Annila 2023). Thus, one way to formulate causal realism while circumventing Russell's objection is to combine the principle of entropy with the symmetric determination given by other (domain specific) laws (Albert 2000).

It is also possible to define causation in such a way that asymmetry is derived not by directly appealing to entropy but through contingent facts about our world that are related to the second law of thermodynamics. This solution is the one proposed by David Lewis (1979), for whom the asymmetry of counterfactual dependence reduces to the asymmetry between small and large miracles, or the asymmetry of overdetermination.[19]

To be clear, the mere possibility of formulating this kind of (*indirect*) causal realism would, theoretically speaking, pose a problem for Russell's first argument. However, the fact that there are *existing positions* employing that possibility – that accept his premises while effectively bypassing his conclusion – explicitly and convincingly demonstrates the serious shortcomings of his argument.

Let us briefly address a potential worry here. Given the sophistication and idiosyncrasies of contemporary causal realism, one might argue that the proclaimed survival of causal realism is merely nominal rather than substantive. This concern stems from the sense that *much has been lost* in the transition from the traditionally conceived notion of causal realism to this current *weak*, *indirect*, or *light* realism.

---

to the fact that although physical laws may not be formally asymmetrical, the processes they describe often are in fact physically asymmetrical" (Lipkind, 1979, p. 719).

[19] Lewis himself emphasizes that this asymmetry is a contingent fact and a local characteristic of our world (see Lewis, 1979, p. 475). However, Adam Elga (2000) argues that even in Lewis' theory, certain restrictions related to entropy must be explicitly added to ensure asymmetry, as he believes that, considering insights from modern quantum physics, Lewis's original solution is not sufficiently robust. For an analysis of the asymmetry of counterfactual dependence within Lewis's theory, see: Đorđević & Ostojić (2015).

While our analysis does involve revising and redefining the term *causal realism*, two points must be emphasized. First, this intervention is not all that uncommon, and we are not alone in this revisionist endeavor; many authors are inclined[20] – against the traditional conception of causal realism that gives rise to these worries – to understand causal realism in terms nearly identical to those we propose.

Second, the basis for our redefinition is not *ad hoc* but instead it aims to capture what is arguably *central* to realism. By only giving up on the claim that causation is *de facto* part of the world's inventory (a part of our best scientific explanation of the world, or otherwise posited as a primitive, irreducible relation) we still insist on holding to all those remaining features of the realist causal relation that justify its realist status in the relevant sense. Abandoning the traditional metaphysical view of causation as a fundamental worldly relation does not entail conceding that it is a mere *pragmatic reconstruction* or that it is *divorced from reality*. On our adopted view, theories cannot qualify as causally realist if they are *projectivist* in any sense, if they treat causal relations as inherently dependent on our cognition, mind, language, or pragmatic considerations. Moreover, the example we have emphasized – David Lewis's counterfactual theory – has been analyzed precisely to highlight how the weakly realist position remains *tied to the world*.[21]

It would also be mistaken to dismiss this as a merely semantic dispute over *what to label realism*. Realism entails certain *ontological commitments*, and these – if we are correct – are fully preserved here. Only the *manner* in which these commitments were traditionally upheld has been rejected. A fitting analogy is a famous story from the history of science regarding the *theory of heat*. After discarding the (proper realist) idea of heat as a *distinct substance* (Lavoisier's *caloric fluid*), physicists did not become antirealists about the nature of heat; rather, they eventually settled for theories defining it as a property grounded in the behavior of elementary particles (as manifestation of their kinetic energy). Such theories in no way imply that temperature is a *projection*, a *pragmatic reconstruction*, or a property *divorced from reality*. So, to make the analogy complete, *heat realism* survived the counterargument (or rather crucial experiments) laid against early (boldly) realistic accounts, to continue to live as indirectly realistic, which is still undoubtedly realistic in any sense *worth wanting* (to put it in Dennett's words).

4.2. Lessons from Russell's second argument

Let us now turn to the second argument that Russell has proposed. In presenting his critique of causal realism, we highlighted that his most compelling argument is likely the one that directly attacks the principle of causality. This second argument (A_2) addresses the dilemma concerning the level of specificity in defining events and thrives on the unacceptable

[20]Some of them are cited in this paper: Menzies (2007), Esfeld (2018); and there are still more of those authors which aren't directly relevant to our endeavors but are more or less in sync with us concerning the understanding of causal realism: Paul and Hall (2013), Schaffer (2015), and others.

[21]And tied to *this* world – Lewis's infamous modal realism is not the requirement for the realist interpretation of his theory. As Esfeld and Deckert (2018, pp. 44–5) put it: "[T]he Humean can go for a sophisticated regularity theory [of causation] in terms of counterfactuals [...]. The decisive issue then is that the truth-value of the counterfactuals expressing causal relations supervenes on the configuration of matter and its change in the actual world. To put it differently, no realism about other possible worlds is required to obtain truth-makers for counterfactuals in Humeanism."

consequences of both potential solutions to this problem. This argument has become a classic issue for all theorists of causation who have, in one way or another, sought to adopt the idea of causation as determination or aim to refute it.[22]

Contrary to the discussion of the previous argument – where the goal was to show how realism could be defended by pointing out its problems – this second argument will be examined primarily from the perspective of the lesson that causal realism draws from its plausibility and persuasiveness. Even more specifically, in the following passages, we will present how Lewis' causal theory (serving as a good example of modest causal realism) avoids the problem highlighted by this Russellian argument.[23]

Lewis's theory of causation (Lewis 1973b, 1986a, 2004), mentioned above more than once, is undoubtedly among the most significant contemporary causal theories. In the context of analyzing Russell's critique of causal realism, two things are particularly noteworthy: 1. this theory unequivocally has a realist character and represents a version of – what we called – light (or minimal, or indirect) causal realism, and 2. the central definition of causation in this theory is surprisingly similar to the principle of causality that Russell so persuasively critiques. With this second point in mind, it is certainly a good question – how has this version of causal realism managed to avoid Russell's second and more compelling argument?

First, let's note that the similarity between Lewis's analysis of causation and the causality principle attacked by Russell (R_{cp}) lies in their shared origin. Both statements are openly inspired by the positive part of David Hume's famous analysis of causation from his *Enquiry Concerning Human Understanding*. In the part where he addresses causation, we have these two prominent definitions:

> (H_{cp1}) . . . [W]e may define a 'cause' to be an event followed by another, where all events similar to the first are followed by events similar to the second. (H_{cp2}) Or in other words where if the first event hadn't occurred the second wouldn't have occurred either. (Hume 2008: 38)

Although Hume himself clearly thinks that H_{cp2} merely paraphrases what is said in the H_{cp1}, that is surely not the case. In fact, the difference turns out to be extremely important. Lewis famously attends to these "other words" (H_{cp2}) in building his theory of causation, while maintaining that the principle given in H_{cp1} is problematic and unfruitful for contemporary causal theories.[24] In other words, Lewis's utilization of Hume's *other words* very effectively addresses Russell's objection that we are examining here. However, Lewis himself does not emphasize (neither here nor elsewhere), how analysis based on H_{cp2} actually avoids those insurmountable problems that trouble theories (e.g. the regularity theory) associated with the first definition (H_{cp1}). On the other hand, we believe that here we have an important lesson with regards to contemporary causal realism that should be elaborated.

[22] In his overview text on the regularity theory of causation, Psillos (2009) writes about this problem and points to significant commentaries on it by John Venn (1889) and Arthur Pap (1952).

[23] In that, we will be roughly following the analysis offered in Jovanović (2020, pp. 8–11).

[24] As an important shortcoming that suggests the irreparability of the regularity theory (built upon H_{cp1}), Lewis points out its failure to effectively distinguish causes from the effects, epiphenomena, and preempted potential causes (Lewis, 1973b, pp. 556–7).

It is important to note two significant differences between the two cited Hume's sentences. First, in the later sentence (H_{cp2}), causation is defined in terms of singular events, as opposed to the first definition, which is based on the relation between classes of events (or between types of events, since the classes are formed based on similarity to a given event). Second, and equally important: in the second sentence, there is a specific dependence between the non-occurrence of singular actual events, while the first gives an implicative statement between the occurrence of events of one type and the occurrence of events of another type.

These two differences are central to avoiding Russell's objection. Following Hume's second definition, Lewis ties causation to counterfactual dependence between actual singular events. In the most rudimental definition, this Lewisian principle of causality would read: an actual event c is the cause of an actual event e iff it is true that had event c not occurred, event e would not have occurred either.[25] Lewis defines events as properties of spatiotemporal regions, whose identity is not (necessarily) maximally specific, but whose specificity can vary in different conversational contexts (Lewis 1986b). Lewis seeks to preserve, within the theory, the practice of ordinary language, where we typically do not speak of events as being extremely fragile. Thanks to this, the theory allows events to be repeatable, i.e. a given event can occur in different places, at different times, or in slightly different ways or with different surroundings. In this way, the theory seems to be heading towards the problems with the lost determination.

The second significant difference: the type of dependence we have here (between events c and e) is not the same kind of determination we have within first definition, and which led to the problems. Instead of nomological dependence, closely connected to H_{cp1} we now have counterfactual dependence, built upon counterfactual conditionals with their respective semantics and further restrictions (Lewis 1973a, 1973b). Counterfactual conditionals are tied to actuality, in the sense that certain circumstances from the actual world are presupposed and held fixed during their evaluation (including natural laws, but also various contingent facts). When we say of a singular actual event c: "had event c not occurred, then ...," we indeed presuppose much of the circumstances in which c actually occurred. And this is precisely what was missing and causing the problem in the formulation of the principle of causality (both R_{cp} and H_{cp1}). Lewis's theory does not claim "whenever c, then (assuming the laws) always e," because this statement is not tied to actuality in an adequate way, and during evaluation, it allows for variation in circumstances, and changes in circumstances can prevent the occurrence of the effect – as already discussed. In contrast, his definition of causation through counterfactual conditionals inherits from these conditionals a certain "freezing" of circumstances and natural laws (in the sense of holding them fixed),[26] which (in an indirect way) saves it from the second horn of Russell's dilemma as given in A_2.

[25] This is, of course, an oversimplification, for the sake of clarity.

[26] It must be noted – with regards to standard semantics for counterfactuals – that there is no "freezing" and "holding fixed" in the *strict sense*, since nothing that is contingent in any sense (and both circumstances and laws are contingent) is fixed and true throughout the *whole modal horizon*. Only necessary truths are true in all possible words, disregarding their *distance* from the actual world. However, having in mind standard semantics for counterfactuals and the role that comparative similarity of worlds plays in it, there is a sense in which in the nearest and (for the evaluation of the given conditionals) relevant possible worlds those circumstances and certain facts *are* being held fixed (Lewis, 1973a, 1979).

Eagle (2007) insightfully captures this from a formal perspective. For him, the central difference between the two definitions of causation must be seen in the fact that counterfactual implication, unlike the material implication tied to the principle of causality (H_{cp1} and R_{cp}), does not allow the *strengthening of the antecedent* as a valid rule of inference. If it is true that a implies b, then it should hold that a and c together also imply b.[27] In contrast, this law does not hold for counterfactual conditionals.[28] Therefore, the fact that with a change in some circumstances we would no longer have (counterfactual) dependence between cause and effect does not oblige us to say that this dependence does not hold between cause and effect.

We can illustrate this by an example. After inserting money into a coffee machine, a person pressed a button, and shortly afterward, the machine made a cup of coffee. Was that button-pressing the cause of the coffee being made? Undoubtedly, it was. But what do the two mentioned analyses say about this? The principle that Russell critiques (R_{cp}) shows the expected weaknesses. Simply put, not all instances of pressing the button are followed by the preparation of coffee. Conditional: If the person presses the button, the machine will make the coffee – is problematic. Pressing the button is insufficient in various circumstances, ranging from cases where there is not enough money in the machine, or the machine is unplugged, or it malfunctions, etc. So, the expected regularity does not seem to hold.

On the other hand, the counterfactual dependence seems to hold. The conditional: had the person not pressed the button, the coffee machine would not have made coffee – should turn out true. We can attempt to formulate an analogous objection to the one given above – that, with altered circumstances, this second statement would also not be true; for example: had the person in question not pressed the button, but someone else had, the machine would still have made coffee. But this objection is not relevant, since in the actual world these circumstances do not exist, and our counterfactual statement and the way we evaluate it do not oblige us to say that this would hold in all circumstances (i.e. does not commit us to accepting strengthening of the antecedent).

Thanks to this, Lewis's position effectively circumvents Russell's dilemma. In that dilemma, we had two alternatives: either we have local and broadly described events but no determination, or we have regularity and determination but instead of typical events we need to take the states of the world at a given moment as causal relata. Both alternatives, as we have seen, led to the refutation of the principle of causality. By redefining the relation between events by means of counterfactual dependence, the counterfactual theory allows us both to have a definition of causation and for it to refer to local events.

[27] Strengthening the antecedent, as a law of classical logic (where we are dealing with material implication), can be expressed in the form of a valid argument: $(a \supset b) \vDash a \wedge c \supset b$. Such an argument, however, is not valid for counterfactual conditionals: $a \succ b \nvDash a \wedge c \succ b$. (In these formal notations we used the symbol "$\succ$ to denote counterfactual implication. The symbols for conjunction and material implication are, respectively: $\wedge$ and $\supset$; furthermore, the double turnstile, regular ($\vDash$) or negated ($\nvDash$), is used to indicate that the formula on the right (respectively) is, or isn't a semantic consequence of the formula, or set of formulas, on the left.)

[28] This holds for standard semantics that is relevant for our paper. However, as a more general claim this would not be true. Indeed, there are authors who defend the thesis that strengthening the antecedent is still a valid pattern of inference even for counterfactual conditionals. For a detailed analysis of this problem, see (Nute & Cross, 2001).

The other previously mentioned theories that we have considered as causal realism successfully addressed this dilemma in various ways. Woodward's manipulability theory (2003), for instance, does so very similarly to Lewis's – mainly through formal properties of counterfactual dependence. Theories relying on probability (e.g. Hitchcock 2018) could easily accommodate the problem of the absence of determinism when viewed from the perspective of necessity, as long as these patterns of regularity were stochastically identifiable. Process theorists (as well as energy transfer theorists and those similar to them), however, had a somewhat different task. Their way of bypassing Russell's objections did not involve tweaking the principle of causality or redefining causal dependence,[29] but was more focused on modeling the causal relation *on the* laws and entities from fundamental science in a way that at least roughly correlates with the philosophical concept of cause.[30] This plethora of options, together with the Lewisian solution presented in detail above, demonstrates that Russell's second argument, initially deemed intractable, can be successfully met in multiple ways.

5. Conclusion

In this paper, we have analyzed Russell's critique of causal realism and explored how this philosophical position has not only survived but also evolved in response to his arguments. Russell's two central arguments were intended to dismantle the traditional metaphysical commitment to causation as an objective feature of the world. However, as we have seen, causal realism has endured, thanks to both the limitations of Russell's arguments and the innovative and insightful responses developed by contemporary causal realists.

The first argument, which questioned the role (and the possibility of grounding) of causation in fundamental physics, left room for a more modest form of causal realism to emerge. Contemporary theories, such as process theories, counterfactual theories, and interventionist approaches and so on, have redefined causation in terms of supervenience on fundamental physical facts, thereby avoiding the direct confrontation with Russell's critique. Also, those theories somehow find the way around the so-called asymmetry problem. This shift toward a more indirect or modest causal realism has allowed the position to withstand Russell's objections, demonstrating that his critique, while insightful, was not fatal.

The second argument, which focused on the conceptual difficulties in defining causal relations – particularly the problem of specifying events and ensuring the regularity required by the principle of causality – has also been addressed by contemporary theories. Russell's diagnosis of the dilemma faced by causal realists, where broadly defined events fail to ensure regularity and highly specific events render the principle of causality vacuous, was astute. However, modern theorists, such as David Lewis, have circumvented this problem by redefining causation in terms of counterfactual dependence. By focusing on singular events

[29] However, their lack of attention to this exact problem (especially when it comes to causal relata) is, according to Eagle, the reason why process theorists must face the same problems that Russell attributes to the principle of causality. See Eagle (2007, p. 162).

[30] As Bernstein notes, these theories have a somehow problematic "success condition" especially with regards to everyday intuitions (2017, pp. 80–81). Their authors tend to describe causal relations in ways that: 1) can diverge from everyday folk intuitions about causation but also 2) do not directly correspond to our best scientific theories. As Bernstein suggests, their specific project of analysis of causation is rather complex since the goals of scientific accuracy and intuitive plausibility may not always align.

and holding fixed the circumstances in which these events occur, Lewis's theory and others like it have provided a way to preserve the reality of causation without falling into the traps Russell identified.

As we have tried to show, the survival of causal realism can be attributed to a combination of Russell's oversights and the resourcefulness of contemporary realists. By redefining what it means for causation to be *real* and by developing theories that are responsive to the challenges posed by Russell's plausible points, causal realism has not only endured but also thrived. As a consequence, causal realism lives to see another day, not as the naive metaphysical doctrine Russell attacked, but as a refined position that continues to play an important role in the philosophy of causation.

Acknowledgments

This paper was prepared as a part of the project *Philosophical research in the digital sphere – public and scientific aspects*, conducted at the University of Niš – Faculty of Philosophy (No. 423/1-3-01). Also, this study was supported by the Ministry of Science, Technological Development and Innovations of the Republic of Serbia (Contract No. 451-03-137/2025-03/ 200165).

References

Albert, D. Z. (2000). *Time and chance*. Harvard University Press.

Annila, A. (2023). Philosophy of thermodynamics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 381*(2252). https://doi.org/10.1098/rsta.2022.0281

Aronson, J. (1971). On the grammar of 'cause'. *Synthese, 22*, 414–430.

Bernstein, S. (2017). Intuitions and the metaphysics of causation. In D. Rose and J. Schaffer (Eds.), *Advances in experimental metaphysics* (pp. 45–65). Bloomsbury Publishing PLC. Retrieved from https://www3.nd.edu/~sbernste/Intuitions_penultimate.pdf

Blanchard, T. (2016). Physics and Causation. *Philosophy Compass, 11*(5), 256–266. https://doi.org/10.1111/phc3.12319

Corry, R. (2006). Causal realism and the laws of nature. *Philosophy of Science, 73*(3), 261–276.

Dowe, P. (2007). Causal processes. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2010 Edition). Retrieved from https://plato.stanford.edu/archives/sum2010/entries/causation-process/

Dowe, P. (2010). Causal Process Theories. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The Oxford Handbook of Causation* (pp. 213–233). Oxford University Press.

Eagle, A. (2007). Pragmatic causation. In H. Price & R. Corry (Eds.), *Causation, physics, and the constitution of reality: Russell's republic revisited* (pp. 191–223). Oxford University Press.

Edgington, D. (2008). Counterfactuals. *Proceedings of the Aristotelian Society, 108* (part 1), 1–21. https://doi.org/10.1111/j.1467-9264.2008.00233.x

Ehring, D. (1997). *Causation and persistence: A theory of causation*. Oxford University Press.

Esfeld, M. (2011). Causal realism. In D. Dieks, W. Gonzalez, S. Hartmann, M. Stöltzner, & M. Weber (Eds.), *Probabilities, laws, and structures* (pp. 435–460). Springer.

Esfeld, M. & Deckert, D.-A. (2018). *A Minimalist Ontology of the Natural World*. Taylor & Francis Ltd.

Fair, D. (1979). Causation and the flow of energy. *Erkenntnis, 14*, 219–250.

Field, H. (2003). Causation in a physical world. In M. Loux & D. Zimmerman (Eds.), *Oxford handbook of metaphysics* (pp. 435–460). Oxford University Press.

Hájek, A. (2014). Probabilities of counterfactuals and counterfactual probabilities. *Journal of Applied Logic, 12*(3), 235–251. https://doi.org/10.1016/j.jal.2013.11.001

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science, 56*(4), 843–887. https://doi.org/10.1093/bjps/axi147

Hitchcock, C. (2005). What Russell got right. In H. Price & R. Corry (Eds.), *Causation, physics, and the constitution of reality: Russell's republic revisited* (pp. 45–65). Oxford University Press.

Hitchcock, C. (2007). How to be a causal pluralist. In P. K. Machamer & G. Wolters (Eds.), *Thinking about causes* (pp. 200–221). University of Pittsburgh Press.

Hitchcock, C. (2018). Probabilistic Causation, *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), Retrieved from https://plato.stanford.edu/archives/spr2021/entries/causation-probabilistic/

Hume, D. (2008). *Enquiry concerning human understanding*. Retrieved from www.earlymoderntexts.com. (Original work published 1748)

Jovanović, M. Z. (2020). *Causation, conditionals and pluralism* [Unpublished Doctoral Dissertation]. https://nardus.mpn.gov.rs/handle/123456789/17796

Lewis, D. (1973a). *Counterfactuals*. Blackwell.

Lewis, D. (1973b). Causation. *Journal of Philosophy, 70*, 556–567.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs, 13*, 455–476.

Lewis, D. (1986a). Events. In *Philosophical papers II* (pp. 241–269). Oxford University Press.

Lewis, D. (1986b). Postscripts to 'Causation'. In *Philosophical papers II* (pp. 172–213). Oxford University Press.

Lewis, D. (2004). Causation as influence. (Extended version). In J. Collins, N. Hall, & L. Paul (Eds.), *Causation and counterfactuals* (pp. 75–106). MIT Press.

Lipkind, D. (1979). Russell on the notion of cause. *Canadian Journal of Philosophy, 9*(4), 701–720.

Loewer, B. (2007). Counterfactuals and the second law. In H. Price & R. Corry (Eds.), *Causation, physics, and the constitution of reality: Russell's republic revisited* (pp. 293–326). Oxford University Press.

Mackie, J. L. (1974). *The cement of the universe*. Oxford University Press.

Menzies, P. (2007). Causation in context. In H. Price & R. Corry (eds.), *Causation, physics, and the constitution of reality: Russell's republic revisited* (pp. 191–223). Oxford University Press.

Nay, A. (2009). Physical causation and difference-making. *Philosophical Studies, 145*(1), 101–120.

Nute, D., & Cross, C. B. (2001). Conditional logic. In D. M. Gabbay & F. Guenthner (Eds.), *Handbook of philosophical logic* (Vol. 4, pp. 1–98). Springer.

Pap, A. (1952). Philosophical analysis, translation schemas, and the regularity theory of causation. *Journal of Philosophy, 49*, 657–666.

Paul, L. A., & Hall, E. J. (2013). *Causation: A User's Guide*. Oxford: Oxford University Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Pearl, J., & Mackenzie, D. (2019). *The Book of Why: The New Science of Cause and Effect*. Penguin Books.

Price, H. (1997). *Time's Arrow and Archimedes' Point*. Oxford University Press.

Psillos, S. (2009). Regularity theories. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The Oxford handbook of causation* (pp. 131–157). Oxford University Press.

Russell, B. (1948). *Human knowledge: Its scope and limits*. Simon and Schuster.

Russell, B. (1959). On the notion of the cause. In *Mysticism and logic* (pp. 174–201). Allen and Unwin LTD.

Schaffer, J. (2015). Grounding in the image of causation. *Philosophical Studies, 173*(1), 49–100. https://doi.org/10.1007/s11098-014-0438-1

Shinod, N. K. (2019). There, Russell was right: On the status of causality in modern science. *Synthese, 196*(11), 4551–4573.

Spirtes, P., Glymour, C., & Scheines, R. (1993). Causation, Prediction, and Search. In *Lecture notes in statistics*. Springer New York. https://doi.org/10.1007/978-1-4612-2748-9

Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (*American Philosophical Quarterly Monographs, 2*, pp. 98–112). Blackwell.

Suppes, P. (1970). *A probabilistic theory of causality*. North-Holland.

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

Venn, J. (1889). *The principles of empirical or inductive logic*. MacMillan.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Đorđević, V., & Ostojić, J. (2015). Transition periods and miracles. *Belgrade Philosophical Annual, 28*, 89–112.

Balkan Journal of Philosophy (2025), vol. 17, issue 2, 165–185

IDIOLECTAL CONCEPTUAL ANALYSIS

Hristo Valchev*

Abstract: We can distinguish between conceptual analysis of terms in the communal language and conceptual analysis of terms in one's own idiolect. The former can be called "communal conceptual analysis", whereas the latter can be called "idiolectal conceptual analysis". The present paper provides a systematic exposition and defence of idiolectal conceptual analysis, considering its possible and actual use in philosophical inquiry, as well as some of the most important criticisms that can be raised against it. The two most influential kinds of criticism that have been raised against conceptual analysis in general are based on semantic externalism and empirical studies of judgements about possible cases respectively. I argue that both of these kinds of criticism apply mainly to communal conceptual analysis and not to idiolectal conceptual analysis. Furthermore, I argue that even though there is some evidence that most well-known and widely discussed instances of using conceptual analysis have been instances of communal conceptual analysis, idiolectal conceptual analysis can also play a significant role in philosophical inquiry. Thus, the paper suggests a shift of focus from communal conceptual analysis to idiolectal conceptual analysis.

Keywords: conceptual analysis; idiolect; meaning; semantic externalism; experimental philosophy.

1. Introduction

The method of cases is a philosophical method that consists in using judgements about possible cases in order to test philosophical theories. It has been employed throughout the history of philosophy going back as far as Plato. Thus, in the dialogue "Theaetetus", Plato considered several possible answers to the question "What is knowledge?", dismissing each of them on the basis of inconsistency with a certain possible case. For example, he argued that knowledge is not simply true belief, because a skilful lawyer could persuade a person to have a belief that is true without this belief actually being knowledge. In a similar way, when considering his grounds for accepting the reality of the physical world, Descartes imagined a possible case in which a demon is deceiving him by making him see and hear things that are not really there. It was not until the middle the twentieth century, however, that the method of cases became really widespread in philosophical inquiry. Since then, it has generated a substantial body of literature in areas as diverse as epistemology, philosophy of language, philosophy of mind and moral philosophy. Popular possible cases that philosophers have considered include cases of epistemic luck, the case about Twin Earth (Putnam, 1975), the case about the Chinese room (Searle, 1980), the case about Mary in the black-and-white room (Jackson, 1982), the various trolley cases, etc.

It must be noted, however, that philosophers have interpreted the method of cases in different ways (see Kipper, 2012, Chapter 1). Thus, Plato argued that it consists in

*Guangdong University of Foreign Studies, Guangzhou, China; hristo_valchev@yahoo.com

investigating a world of ideas by remembering the experience we had in that world when our souls were there before we were born. Another possible interpretation states that one's judgements about possible cases are based on one's faculty of rational intuition (BonJour, 1998). Adherents of naturalism, on the other hand, argue that one's judgements about possible cases are empirically justified. Thus, Kornblith (1998) argues that they are based on empirical background information, and Williamson (2007, Chapter 6) argues that they are based on our everyday ability to evaluate subjunctive conditionals. More recently, Strevens (2019) has argued that one's judgements about possible cases are based on ordinary beliefs. Probably the most influential interpretation of the method of cases, however, has been the one according to which one's judgements about possible cases are judgements about the application of terms. These judgements can be considered to be intuitive, but they are not based on a separate faculty of rational intuition. Rather, they are intuitions that are acquired as a part of learning the given language and are based on one's semantic competence. If this interpretation is adopted, then the method of cases is called "conceptual analysis".[1]

It was indeed as conceptual analysis that the method of cases initially became commonly used in philosophical inquiry, as a movement centred around usage of the method started to develop within the Oxford school of ordinary language philosophy and soon spread to many places outside Oxford. During the second half of the twentieth century, however, enthusiasm for conceptual analysis started to temper. The method of cases continued to be widely used, but it was not explicitly interpreted as conceptual analysis anymore. We can say that the philosophical practice remained the same, but the way this practice was understood and theorized changed. I think this change occurred mainly due to the rise of naturalism. Influential arguments put forward by Quine (1951, 1960), Putnam (1975) and others convinced many philosophers that there is no a priori knowledge, and since conceptual analysis was supposed to provide a priori knowledge, it became less attractive. We can say that the most influential criticism raised against conceptual analysis from the point of view of naturalism was based on the semantic externalism of Kripke (1980) and Putnam (1975).[2]

The last quarter of a century, however, saw a resurgence of interest in conceptual analysis. This resurgence of interest was initiated by Jackson (1998), but many others began to defend the method as well, including Chalmers (Chalmers & Jackson, 2001); Henderson and Horgan (2011); Kipper (2012); Finlay (2014), etc. The most common way in which these philosophers dealt with the criticism from semantic externalism was by adopting two-dimensional semantics (see Chalmers 2006). However, whether their defence has been successful is still a subject of debate (see Schroeter, 2021). Meanwhile, experimental philosophers raised another influential criticism. It was directed against both conceptual analysis in particular and the method of cases in general and was based on empirical research which

[1]Thus, the term "conceptual analysis" can be understood in two ways. On the one hand, it can be understood as referring to an interpretation of the method of cases, while on the other hand, it can be understood as referring to a separate method in its own right.

[2]Another criticism that can be raised against conceptual analysis from the point of view of naturalism is based on Quine's criticism of the notion of analyticity. As far as the results from conceptual analysis are analytic sentences, undermining the notion of analyticity would also undermine conceptual analysis. Even though this criticism was very influential during the second half of the twentieth century, it seems to have lost some of its appeal. At least, it is rarely discussed nowadays in the context of conceptual analysis. Thus, I have decided not to address it in the present inquiry.

suggests that our judgements about possible cases vary among people from different social groups and are affected by seemingly irrelevant contextual factors (see Knobe & Nichols, 2017). The most common way to deal with this criticism has been to argue that philosophers have better judgements about possible cases than the average person – an argument known as "the expertise defence." However, the expertise defence has been demonstrated rather conclusively to be false empirically (see Knobe & Nichols, 2017).

In the present paper, I argue that both the criticism based on semantic externalism and the criticism based on studies in experimental philosophy apply mainly to conceptual analysis of terms in the communal language and not to conceptual analysis of terms in one's own idiolect. For convenience, we can call these two kinds of conceptual analysis "communal conceptual analysis" and "idiolectal conceptual analysis" respectively. Thus, I argue that the said kinds of criticism apply mainly to communal conceptual analysis and not to idiolectal conceptual analysis. Furthermore, I argue that even though most well-known and widely discussed instances of using conceptual analysis have been instances of communal conceptual analysis, idiolectal conceptual analysis can also play a significant role in philosophical inquiry. Thus, we can say that the paper provides a defence of idiolectal conceptual analysis. The text is divided into five sections. In the second section, I provide more precise definitions of communal conceptual analysis and idiolectal conceptual analysis; the third section is concerned with the use of idiolectal conceptual analysis in philosophical inquiry, addressing both its possible and its actual use; the fourth section is concerned with the possible criticisms of idiolectal conceptual analysis; and the fifth section contains some concluding remarks.

2. What Is Idiolectal Conceptual Analysis?

2.1. What Is Conceptual Analysis?

The term "conceptual analysis" was introduced in analytic philosophy by Grice (1958), who used it in order to describe the method employed by the Oxford school of ordinary language philosophy. He wrote the following.

> To be looking for a conceptual analysis of a linguistic expression E is to be in a position to apply or withhold E in particular cases, but to be looking for a general characterization of the types of case in which one would apply E rather than withhold it. And we may notice that in reaching one's conceptual analysis of E, one makes use of one's ability to apply and withhold E, for the characteristic procedure is to think up a possible general characterization of one's use of E and then to test it by trying to find or imagine a particular situation which fits the suggested characterization and yet would *not* be a situation in which one would apply E. If one fails, after careful consideration on these lines, to find any such situation, then one is more or less confident that the suggested characterization of the use of E is satisfactory. But one could not test a suggested characterization in this way, unless one relied on one's ability to apply or withhold E in *particular* cases. (Grice, 1958, p. 174)

Even though Grice is talking in this passage about the application of linguistic expressions, it is evident from the mere term "conceptual analysis", as well as from some of his

remarks, that he is actually interested in the application of the concept that the given linguistic expression expresses. Two things need to be noted here. First, not all linguistic expressions express concepts. We usually call the linguistic expressions that do express concepts "predicates" or simply "terms." Second, the application conditions of a term are not necessarily the same as the application conditions of the concept that the term expresses. This can be demonstrated through a version of an example by Glock (Glock, 2011, pp. 159–160). Imagine that the prime minister is giving a speech in the House of Commons and is saying something which is not true. The members of Parliament cannot say that he is lying, because this would be too impolite – they need to use a more appropriate expression such as "being economical with the truth". In this sense, the term "lie" is not applicable to what he is saying, even though the concept of a lie (the concept that the term "lie" expresses) is. At the time of writing the cited passage, Grice might not have been aware of the distinction between strictly conceptual matters and such pragmatic matters as regards the usage of terms, but in his later work, he paid special attention to drawing this distinction (see Grice, 1989).

In the context of this distinction, we can use the concept of semantic application of terms (see Valchev, 2018). We can say that a term is semantically applicable to an object, if and only if the object falls under the concept that the term expresses. Thus, the judgements that one makes when conducting conceptual analysis are judgements about the semantic application of terms in possible cases. Accordingly, the ability to apply or withhold E in particular cases, which Grice is talking about in the cited passage, is one's semantic competence. Furthermore, we can say that the conditions that are necessary and sufficient for an object to satisfy in order for a term to be semantically applicable to it are the semantic application conditions of the term. Thus, the general characterization of the types of case in which one would apply E rather than withhold it, which Grice is referring to as the goal of conceptual analysis, can be understood as a conclusion about the semantic application conditions of the term. In this sense, conducting conceptual analysis consists in using judgements about the semantic application of a term in possible cases as evidence for a conclusion about the semantic application conditions of the term. As far as to know the semantic application conditions of a term is the same as to know its meaning, we can also say that the meaning of a term is the set of its semantic application conditions. In this sense, the goal of conceptual analysis is to draw a conclusion about the meaning of the given term.

It is not entirely clear whether Grice thought that the aim of the analysis should be to provide conditions for the semantic application of the given term that are both necessary and sufficient[3], but we can assume that any investigation in which consideration of possible cases is used as a basis for a conclusion about the semantic application conditions of a term consists in conceptual analysis – regardless of whether the conclusion is that the given conditions are only necessary, only sufficient, both necessary and sufficient or neither necessary nor sufficient.[4] This is the way the method is understood by Kipper (2012, p. 252), Henderson

[3]In the passage cited above, Grice seems to be talking about conditions that are only sufficient, but he also says that conceptual analysis is supposed to answer questions of the kind "What is X?" (Grice, 1958, pp. 173–4), and such questions seem to ask about semantic application conditions that are both necessary and sufficient.

[4]See Valchev (2018). In this way, the scope of the method is made broader and the criticism of the method which points to the fact that there have been very few complete analyses in the sense of analyses that provide

and Horgan (2011, p. 39) and Chalmers and Jackson (2001, p. 322) among others. It should also be noted that philosophers have also drawn conclusions about the semantic application conditions of terms on the basis of empirical studies that look into how the given term is used in possible cases. Such studies can also be interpreted as consisting in conceptual analysis (see Valchev, 2018, 2022). This interpretation is adopted for example by Glasgow (2008, p. 333), Kriegel (2017, p. 30), Overton (2013, p. 1383), Sytsma (2010, p. 427), Tallant and Andow (2020, p. 6) and Zach (2019). If the interpretation in question is adopted, then we can call the conceptual analysis in which one sits in one's armchair and imagines possible cases "armchair conceptual analysis", and the conceptual analysis that is based on empirical research – "empirical conceptual analysis" (Valchev, 2022). It must be noted, however, that the criticisms of conceptual analysis addressed in the fourth section of the text are directed against armchair conceptual analysis in particular.

2.2. Idiolectal Conceptual Analysis and Communal Conceptual Analysis

Consider the following well-known possible case introduced by Burge (1979). John has pain in the thigh and goes to the doctor saying that he has arthritis in the thigh, to which the doctor replies that it cannot be arthritis, since arthritis is specifically an inflammation of joints. John does want to conform to the linguistic usage of the community, so he starts to use the term "arthritis" in accordance with the doctor's explanation. This possible case can be interpreted in two different ways. One possible interpretation would be to say that the term "arthritis" in John's idiolect referred to an inflammation of joints even before John's usage of the term was corrected by the doctor, whereas another possible interpretation would be to say that the meaning of the term in John's idiolect changed. These two interpretations correspond to two different views of meaning, which can be described as "anti-individualism" and "individualism" respectively (see Wikfross, 2017). According to anti-individualism, the meaning of a term in the idiolect of a particular speaker is determined by the meaning of the term in the communal language (Burge, 1979; Williamson, 2007), whereas according individualism, idiolects are independent from communal languages (Davidson, 1987; Farkas, 2008). For the purposes of the present inquiry, I have adopted individualism.

Individualism about meaning entails a view of language that can be elaborated in the following way (see Horwich, 2020). The meaning that a given speaker ascribes to a given linguistic expression is the idiolectal meaning of that expression for that speaker. All pairs of a linguistic expression and its meaning that a speaker employs in the production and comprehension of speech constitute the idiolect of the speaker. A group of speakers with similar idiolects constitute a linguistic community. The meaning that a linguistic community as a whole ascribes to a given linguistic expression is the communal meaning of that expression for that linguistic community. All pairs of a linguistic expression and its meaning that a linguistic community as a whole employs in the production and comprehension of speech constitute the communal language of the linguistic community. There are two ways in which the communal meaning of a linguistic expression is determined. On some occasions, the communal meaning of a linguistic expression is the meaning that the vast majority of speakers ascribe to it, whereas on other occasions, the communal meaning of a linguistic expression

semantic application conditions that are both necessary and sufficient (see Williamson, 2000, p. 31; Fodor, 1981, p. 284) is in some sense avoided.

is determined by a certain group of experts within the community (scientists, doctors, etc.), who have been granted this special responsibility (and the authority which comes with it).

Even though the distinction between idiolects and communal languages has sometimes been discussed in the context of conceptual analysis (see Cekiera, 2023; Chalmers, 2006, 2011; Glock, 2017; Jackson, 2011; Kölbel, 2021), few authors have stated specific interest in investigating the communal or the idiolectal meaning of terms. One exception is Grice, who explicitly stated that he was interested in conducting conceptual analysis of terms in his own idiolect (Grice, 1958, p. 175). Austin (another prominent member of the Oxford school, who Grice called "the Master" of this kind of philosophy – see Grice, 1989, p. 182) did not address the matter explicitly, but included in his description of the method (Austin, 1966, p. 429) a requirement that we reach an agreement as regards the semantic application of the given term, which suggests interest in the communal meaning. More recently, Grindrod (2023) has also stated interest in the communal meaning. We can call the conceptual analysis in which one uses judgements about the semantic application of a term in possible cases as evidence for a conclusion about the communal meaning of the term "communal conceptual analysis", and the conceptual analysis in which one uses judgements about the semantic application of a term in possible cases as evidence for a conclusion about the meaning the term has in the idiolect of a particular speaker – "idiolectal conceptual analysis".

It must be noted that both idiolectal conceptual analysis and communal conceptual analysis can be conducted either from the armchair or empirically. In general, one's judgements about the semantic application of a term in possible cases provide one with evidence for a conclusion about the meaning of the term in one's own idiolect, but one is usually justified in assuming that the meaning one ascribes to the term is the same as the meaning that the linguistic community as a whole ascribes to the term. Under this assumption, one's judgements about the semantic application of a term in possible cases can also provide one with evidence for a conclusion about the communal meaning of the term, although this evidence would not be as good as the evidence for a conclusion about the meaning of the term in one's own idiolect. Empirical methods like questionnaire surveys and corpus analysis, on the other hand, are normally used in order to gather evidence about the communal meaning of terms, but they can also provide evidence for a conclusion about the meaning of a term in the idiolect of a particular speaker. One can argue that in general empirical conceptual analysis provides better justification for a conclusion about the meaning of a term in the communal language than armchair conceptual analysis, but if one is interested in the meaning of a term in one's own idiolect, then armchair conceptual analysis seems to be a much easier and more reliable option (see Valchev, 2023).

3. The Use of Idiolectal Conceptual Analysis in Philosophical Inquiry

3.1. Possible Use of Idiolectal Conceptual Analysis

If one is interested in doing research in areas like history of philosophy or sociology of philosophy, then one can use idiolectal conceptual analysis in order to investigate the meanings of terms in the idiolects of philosophers other than oneself. Thus, one might use idiolectal conceptual analysis in order to investigate for example the meaning of the term "absolute spirit" in the idiolect of the German philosopher Hegel. If, however, one

is interested in doing research in philosophy *per se*, then the usage of idiolectal conceptual analysis is generally limited to investigating the meaning of terms in one's own idiolect. Furthermore, as noted in the previous section, if one is interested in the meaning of a term in one's own idiolect, then conducting conceptual analysis from the armchair seems to be a much easier and more reliable option than conducting conceptual analysis empirically. Thus, we can say that the use of idiolectal conceptual analysis in philosophical inquiry would normally consist in conducting armchair conceptual analysis of terms in one's own idiolect.

It must be noted, however, that in order to conduct armchair conceptual analysis of a term in one's own idiolect, one needs to make judgements about the semantic application of the term in possible cases, and thus, needs to know the meaning of the term already. This is the so-called "paradox of analysis". This paradox can be resolved by making a distinction between two kinds of knowledge of the meaning of a term – procedural knowledge and propositional knowledge. One's procedural knowledge of the meaning of a term is roughly the same as what I previously called "semantic competence", and consists in the ability to distinguish between instances in which the term is semantically applicable and instances in which the term is not semantically applicable, whereas one's propositional knowledge of the meaning of a term consists in a justified true belief that the term has certain semantic application conditions.[5]

Thus, if one is able to tell that the term "bachelor" is semantically applicable to a friend called Victor, but not to a friend called George, then one has procedural knowledge of the meaning of the term, whereas if one has a justified true belief that being unmarried is a necessary condition for the semantic application of the term "bachelor", one has propositional knowledge of the meaning of the term. Regarding the meaning of a term in one's own idiolect, one has procedural knowledge, but might not have propositional knowledge. In this case, one can use idiolectal conceptual analysis in order to acquire such knowledge.

On the one hand, one might be interested in acquiring propositional knowledge of the meaning a term "ordinarily" has in one's idiolect – the ordinary meaning of terms like "knowledge", "freedom", etc. can itself be a subject of philosophical interest. While engaging in philosophical inquiry, however, one might want to use concepts that are not expressed properly by any of the terms in one's idiolect as considered in their ordinary meanings. In this kind of situation, one needs to ascribe to some term a meaning in which it expresses the given concept. The term might have or might have not existed previously in one's idiolect. In the latter case, we have an instance of introducing a new term into one's idiolect, whereas in the former case, we have an instance of changing the idiolectal meaning of a term, at least in the context of one's philosophical inquiry.[6] Introducing a new term into one's idiolect is normally an intentional and conscious process, but the ascription of a new idiolectal meaning to an old term might be done unintentionally and unknowingly. If one is unaware of the fact

[5]This distinction is similar to the distinction Moore (1903) makes between knowing the meaning of a term and knowing its analysis, as well as to the distinction Ryle (1949) makes between knowledge how and knowledge that. I have used the traditional definition of knowledge as justified true belief in order to define one's propositional knowledge of the meaning of a term, since in this context, it seems to be immune to the type of counterexamples it is normally susceptible to.

[6]Naturally, the same term can have different meanings in different contexts. Accordingly, one can use conceptual analysis in order to investigate the idiolectal meaning of a term in a particular context, i.e. what Grice (1989) calls "applied timeless-idiolectal meaning".

that one has changed the idiolectal meaning of a term, then one has procedural knowledge of the meaning of the term, but not propositional knowledge. Accordingly, one can use idiolectal conceptual analysis in order to acquire propositional knowledge.[7] Thus, as far as using a term in a meaning which is different from its ordinary meaning is the same as using the term in a "technical" meaning, we can say that idiolectal conceptual analysis can be used in philosophical inquiry in order to acquire propositional knowledge of both ordinary meanings and technical meanings of terms.

Regardless of whether the meaning ascribed to a philosophical term in one's idiolect is ordinary or technical, however, acquiring propositional knowledge of it can help one explicate one's own philosophical theories and arguments and thus understand them better, find potential faults in them, etc. Besides oneself, the results from the analysis would also be interesting and useful to anyone else who ascribes to the given term the same idiolectal meaning but does not possess propositional knowledge of this meaning. It needs to be noted, however, that the results from the analysis might also be of interest to people who ascribe to the given term a different meaning. The fact that they ascribe to the term a different meaning would normally not hinder their understanding, and actually, the results from the analysis would *aid* their understanding. On the one hand, such people might just be interested in learning about one's point of view, but on the other hand, they might also find the results from the analysis useful. For one, they might be using a different term in order to express the same concept. And even if they do not possess the given concept, the results from the analysis would help them acquire it. As far as they find this concept to be useful, they will also find the results from the analysis to be useful.

Furthermore, acquiring propositional knowledge of the idiolectal meaning of philosophical terms can help resolve conceptual disagreements between philosophers, i.e. disagreements that are caused by the fact that the two sides ascribe to the same term two different meanings. In order to resolve such a disagreement, we need to demonstrate that the two sides ascribe to the term two different meanings. This can be done using conceptual analysis and if one or both sides are interested in the meaning of the term in some idiolect, then the conceptual analysis in question would be idiolectal conceptual analysis. If the two sides

[7] As far as changing the meaning of a term entails a revision of the corresponding concept, we can say that using idiolectal conceptual analysis in order to analyse a term whose meaning one has unintentionally changed while engaging in philosophical inquiry results in an analysis of a revised concept. However, the revision itself is not a part of the usage of conceptual analysis. In this relation, it can be noted that some authors do understand conceptual analysis as involving not only describing but also *pre*scribing the application conditions of the given concept, thus engaging in concept revision (Kölbel, 2021; Sekowski, 2022; Stuart, 2015). Concept revision might well play an important role in philosophical inquiry, but it does not fall under the concept expressed by the term "conceptual analysis" in the meaning it normally has in analytic philosophy (see Valchev 2018). One might think that if we extend the meaning of this term in such a way that conceptual analysis does involve concept revision, then the concept of conceptual analysis would be improved, but actually if such a meaning change is made, the new concept of conceptual analysis would be too different from the old concept in order for the change to constitute an improvement of the old concept (Valchev 2018, p. 139). We would not improve the old concept but simply introduce a new one. Thus, it is not clear why these authors use the term "conceptual analysis" in the said way. One possible explanation is that they have made an unsuccessful attempt to explain the fact that the usage of conceptual analysis might result in an analysis of a revised concept. The exact relationship between their views and idiolectal conceptual analysis, however, lies beyond the scope of the present paper.

are interested in the meaning of the term in different languages – for example, if one side is interested in the meaning of the term in their own idiolect, whereas the other side is interested in the meaning of the term in some communal language – then there aren't any further steps that can be made towards resolving the disagreement. Neither of the two sides are "right" or "wrong" – they are just talking about different things. However, we can say that in such a situation, the mere demonstration of the fact that the two sides ascribe to the term two different meanings consists in a kind of resolution of the disagreement. Thus, if the said demonstration is done using idiolectal conceptual analysis, then idiolectal conceptual analysis is used as a way to resolve the disagreement.

If, on the other hand, the two sides are interested in the meaning of the term in the same language, then the mere demonstration of the fact that they ascribe to the term two different meanings is not sufficient for resolving the disagreement. In this case, one can use conceptual analysis in order to investigate the meaning of the term in the given language. If the results from the analysis suggest that the meaning of the term in the language in question is the same as the meaning ascribed to the term by any of the two sides in the disagreement, then it can be assumed that the side in question is "right", and if the results from the analysis suggest that the meaning of the term in the language in question is different from the meaning ascribed to the term by any of the two sides in the disagreement, then it can be assumed that the side in question is "wrong." Thus, the disagreement would be resolved (see Sytsma, 2010; Valchev, 2022). However, conceptual disagreements in which the two sides are interested in the meaning of the given term in the same idiolect seem very unlikely to appear when engaging in philosophy *per se* as opposed to history of philosophy. As I noted above, the usage of idiolectal conceptual analysis in philosophical inquiry is generally limited to investigating the meaning of terms in one's own idiolect. Thus, we can say that idiolectal conceptual analysis has practically no use in philosophical inquiry as a way to resolve conceptual disagreements by demonstrating that either of the two sides in the disagreement are right or wrong. In this sense, idiolectal conceptual analysis has practically no use in philosophical inquiry for criticising the views of other philosophers or defending one's own views from criticism.

3.2. Actual Use of Idiolectal Conceptual Analysis

As I noted earlier, there are few explicit proponents of idiolectal conceptual analysis or communal conceptual analysis. Thus, if we assume that drawing a conclusion about the meaning of a term in a certain language is necessarily a conscious mental process, then we need to admit that there are few instances of both idiolectal conceptual analysis and communal conceptual analysis. If, on the other hand, we assume that it is possible for one to draw a conclusion about the meaning of a term in a certain language without being aware of the fact that the conclusion is about the meaning of the term in precisely that language, then we can look at instances in which a philosopher has used conceptual analysis without stating specific interest in the communal or the idiolectal meaning of the given term, and try to interpret them as being instances of idiolectal conceptual analysis or communal conceptual analysis. If we also assume that all instances of using the method of cases are instances of using conceptual analysis, then we can try to interpret any instance of using the method of cases as being an instance of idiolectal conceptual analysis or communal conceptual analysis.

This kind of interpretation can be based on different considerations. In the previous section, I argued that idiolectal conceptual analysis has practically no use in philosophical inquiry for criticising the views of other philosophers or defending one's own views from criticism. Thus, if conceptual analysis is used for such purposes, then it can be assumed that its usage constitutes an instance of communal conceptual analysis. It must be noted, however, that it would sometimes be unclear whether conceptual analysis is used for such purposes. We can say that the evidence that it is used for such purposes is evidence that its usage constitutes an instance of communal conceptual analysis. Another kind of evidence we can take into account when trying to identify instances of idiolectal conceptual analysis and communal conceptual analysis is provided by the linguistic expressions the given philosopher is using when describing judgements about possible cases. Expressions that usually refer to the whole linguistic community such as "we" or "our intuitions" would be evidence that the philosopher is conducting communal conceptual analysis, whereas expressions that usually refer only to oneself such as "I" or "my intuition" would be evidence that the philosopher is conducting idiolectal conceptual analysis.[8] We can say that on the whole, these two kinds of evidence suggest that most well-known and widely discussed instances of using the method of cases are actually instances of communal conceptual analysis.

Consider the following examples taken from some of the most well-known works of the philosophers from the Oxford school of ordinary language philosophy. When Ryle employed conceptual analysis in his 1949 book *The Concept of Mind*, his purpose was to criticise a certain view about mental concepts held by Descartes and others (calling it "the myth of the ghost in the machine"). Accordingly, he used expressions that refer to the whole linguistic community, talking for example about the way "we" use the terms "voluntary" and "involuntary" (Ryle, 1949, p. 56). In a similar way, when Strawson employed conceptual analysis in his 1950 paper "On Referring", his purpose was to criticise Russell's theory of definite descriptions, and he used expressions that refer to the whole linguistic community, saying for example that if the sentence "The king of France is wise" was uttered now, "no one" would say that it is meaningless (Strawson, 1950, p. 321). Thus, even though Ryle and Strawson did not address the distinction between idiolects and communal languages explicitly, we can assume that they were both conducting communal conceptual analysis. In contrast, when Grice was conducting conceptual analysis of the concept of meaning, his purpose was not to criticise the views of any other philosophers, and he often used the personal pronoun "I", saying for example: "I find I want to deny that [the term "means" is semantically applicable in this-and-this possible case]" (Grice, 1950, pp. 382–383).

Actually, Grice seems to be the only member of the Oxford school who was interested in conducting idiolectal conceptual analysis. This is consistent with the fact that unlike other members of the Oxford school and ordinary language philosophers in general, he was not an enthusiastic proponent of the argument according to which philosophical problems are rooted in deviations from the ordinary use of language. As far as such deviations are demonstrated through conducting conceptual analysis, we can say that this argument entails using conceptual analysis in order to criticise the views of philosophers who deviate from the

[8]Note that this kind of evidence is not infallible. For example, one might be using the pronoun "we" for stylistic reasons or in order to refer to oneself and anyone else who ascribes to the given term the same meaning as oneself.

ordinary use of language, and therefore, implies interest in conducting communal conceptual analysis.[9] Unsurprisingly, the late Wittgenstein – who was one of the originators of this argument and a major source of influence for the Oxford school – often used expressions that refer to the whole linguistic community when describing judgements about possible cases. Thus, he famously stated that "for a large class of cases – though not for all – in which *we* employ the word "meaning," it can be defined thus: the meaning of a word is its use in the language" (Wittgenstein, 1951, §43; added emphasis).

Outside of ordinary language philosophy, conceptual analysis has also played a central role in the Canberra plan (see Braddon-Mitchell & Nola, 2009). On the one hand, the term "the Canberra plan" can be understood as referring to a philosophical movement, whereas on the other hand, it can be understood as referring to a philosophical method for reducing matters described in one vocabulary to matters described in another vocabulary (e.g. mental states to brain states) that the participants in the movement endorse. This method consists of two steps: 1) collecting "platitudes" about some X in order to define the role that a term which refers to X plays in the theory it is a part of; and 2) finding something in the world which plays this role (Nolan, 2009). As far as the platitudes in question can be understood as possible cases of language use and defining the role of the given term can be understood as defining its semantic application conditions, we can say that the first step consists in conducting conceptual analysis. As far as the goal of the Canberra plan is to reduce matters described in one vocabulary to matters described in another vocabulary, we can say that its usage of conceptual analysis does not include criticising the views of other philosophers or defending one's own views from criticism. However, the proponents of the Canberra plan have been more interested in defending conceptual analysis from the criticisms raised against it rather than using conceptual analysis in order to define the roles of terms in accordance with the first step described above and carry out the corresponding reduction, as evident from the fact that they have provided very few actual analyses.[10]

The method of cases has also been used extensively in the debate about the definition of knowledge (see Ichikawa & Steup, 2018). The philosophers who take part in this debate discard previously proposed definitions by describing possible cases that constitute counterexamples to these definitions and attempt to formulate new definitions that avoid these counterexamples. As far as one's discarding of a previously proposed definition implies criticising the view of the philosopher who has proposed the definition, we can say that one is conducting communal conceptual analysis. Discarding a given definition does not necessarily imply criticising the given philosopher's view – one might be discarding a definition because it does not fit one's own concept of knowledge, etc. However, the philosophers who take

[9]We can talk about the ordinary meaning a term has in the idiolect of a particular speaker, as well as the ordinary meaning a term has in the communal language. In the context of the said argument, the expression "ordinary meaning" seems to refer to the latter.

[10]In this relation, Lewis states that he offers not analysis but a recipe for analyses (Lewis, 1994, p. 421). It can also be noted that in the few cases in which the proponents of the Canberra plan have provided actual analyses, they tend to use linguistic expressions that refer to the whole linguistic community when describing judgements about possible cases. For example, in his analysis of the concept of colour, Jackson says that "we" do not use "red" as the name of the experience itself, but rather of the property of the object putatively experienced when it looks red (Jackson, 1998, p. 89). Thus, we can say that there is actually some (rather weak) evidence that the proponents of the Canberra plan have also been more interested in conducting communal conceptual analysis rather than idiolectal conceptual analysis.

part in the debate about the definition of knowledge often use linguistic expressions that do seem to imply criticism. For example, when Gettier discarded the traditional definition of knowledge as justified true belief, he said that it "is false" or "fails" (Gettier, 1963, p. 121); discarding a definition proposed by Unger, Goldman argued that it is "not very satisfying" (Goldman, 1976, p. 773); discarding a definition proposed by Sosa, Comesaña argued that the safety requirement it posits is "incorrect" (Comesaña, 2005, p. 397), etc. Thus, we can say that there is some evidence that the participants in this debate have also been more interested in conducting communal conceptual analysis rather than idiolectal conceptual analysis.

We can also take into account the way Ayer and Chisholm used conceptual analysis in order to formulate the traditional definition of knowledge as justified true belief. Formulating this definition, they were not criticising any philosophical views regarding knowledge, but they used linguistic expressions that refer to the whole linguistic community, talking for example about the way "we" use the term "knows" (Ayer, 1956, p. 26; Chisholm, 1957, p. 16). Another particular instance of using the method of cases that has received a lot of scholarly attention and has often been interpreted as an instance of conceptual analysis (e.g. in Jackson, 1998) is Putnam's argument based on the possible case about Twin-Earth (Putnam, 1975). On the one hand, the argument was directed against descriptivist theories of meaning, whereas on the other hand, Putnam used expressions that refer to the whole linguistic community in order to describe judgements about possible cases, talking for example about what "an English speaker" would have called "water" in 1750 (Putnam, 1975, p. 142). Thus, there is some evidence that Ayer, Chisholm and Putnam have also been more interested in conducting communal conceptual analysis rather than armchair conceptual analysis.

4. Possible Criticisms of Idiolectal Conceptual Analysis

4.1. Semantic Externalism

Semantic externalism is the philosophical view according to which linguistic meaning depends on factors that are "external" to the speakers of the given language. This view has its roots in the work of Kripke and Putnam (Kripke, 1980; Putnam, 1975), who argued against the theories of meaning that were popular amongst analytic philosophers at that time – most prominently, the "descriptivist" theories, put forward by Frege and Russell (see Frege, 1982; Russell, 1905). Kripke and Putnam did not explicitly put forward a criticism against conceptual analysis based on semantic externalism, but such a criticism has been put forward by other philosophers including Devitt and Sterelny (1999), Laurence and Margolis (2003), Haslanger (2012) and Fisher (2015) among others. Haslanger formulates the criticism in the following way.

> In undertaking conceptual analysis of, say, Fness (in our case, Fness might be 'Blackness,' 'Whiteness,' 'Asianness,' or the broader category, 'race'), it is typically assumed that it is enough to ask competent users of English under what conditions someone is F. After all, if competent speakers know the meaning of their terms, then all that is needed is linguistic competence to analyze

> them. However, this stance is not plausible if one takes into account arguments in philosophy of language over the past 30 years that call into question the assumption that competent users of a term have full knowledge of what the term means. This assumption in particular is challenged by the tradition of semantic externalism. Externalists maintain that the content of what we think and mean is determined not simply by what we think or intend, but at least in part by facts about our social and natural environment. For example, one can be competent in using the term 'water' without knowing that water is H_2O; one can use the term 'elm' meaningfully even if one cannot tell the difference between a beech and an elm. When I say, 'Elm trees are deciduous' I say something meaningful and true, even though I couldn't identify an elm or give any clear description of one. (Haslanger, 2012, p. 304)

Since Haslanger is criticising conceptual analysis in the context of discussing race, she is talking about analysing concepts like whiteness and blackness, but what she is saying applies to conceptual analysis in general. In short, the criticism consists in the claim that arguments in the tradition of semantic externalism have challenged an assumption made by the proponents of conceptual analysis – namely, that competent users of a term have full knowledge of what the term means. The arguments in question point to two features of language, illustrated in the two examples given by Haslanger in the cited passage – "reference magnetism" and "division of linguistic labor" (Haslanger, 2012, p. 305). Reference magnetism is characteristic of terms whose extension is determined by a relation of sameness to a certain paradigm, which acts as "a reference magnet" (Putnam, 1973, 1975; Kripke, 1980). For example, if we assume that the reference magnet of the term "water" is the liquid in our rivers and seas, we can say that the term refers to anything which is the same as the liquid in our rivers and seas. Thus, imagine a planet called Twin-Earth on which everything is the same as on Earth except the fact that the microstructure of the liquid in the rivers and seas is not H_2O but XYZ. This liquid might have all the typical characteristics of water, but it would not be the same as the liquid in our rivers and seas, and therefore, it would not be water, i.e. it would not be a part of the extension of the term "water". Division of linguistic labour, on the other hand, consists in the fact that certain members of the linguistic community are responsible for the meaning of certain terms (Putnam, 1975; Burge, 1979). Thus, botanists are responsible for the meaning of terms like "beech" and "elm", so most other people cannot distinguish between the extensions of these terms.

The references to linguistic competence and knowledge of the meaning of terms that Haslanger makes in the cited passage can be understood as references to one's procedural knowledge of linguistic meaning. Thus, Haslanger argues that one cannot conduct conceptual analysis, because one lacks full procedural knowledge of linguistic meaning. We can say that to have full procedural knowledge of the meaning of a term is to be able to tell for each sufficiently specified possible case whether the term is semantically applicable in it. In this sense, the argument can also be formulated as follows: one cannot conduct conceptual analysis of a given term, because one is unable to tell for each sufficiently specified possible case whether the term is semantically applicable in it. Thus, reference magnetism and division of linguistic labour can be understood as two possible reasons why one might be unable to tell for each possible case whether a given term is semantically applicable in it.

These two reasons can be interpreted as corresponding to two different arguments. These arguments can be formulated in the following way.[11]

Argument from Reference Magnetism

(1) A term characterized by Reference Magnetism is semantically applicable in a sufficiently specified possible case, iff an object from the case is the same as the corresponding reference magnet.

(2) In order to be able to tell whether an object is the same as a reference magnet, one needs to know the specifics of the essence of the reference magnet.

(3) In order to be able to tell whether a term characterized by Reference Magnetism is semantically applicable in a sufficiently specified possible case, one needs to know the specifics of the essence of the corresponding reference magnet. – from (1) and (2)

(4) If one does not know the specifics of the essence of the corresponding reference magnet, one is not able to tell whether a term characterized by Reference Magnetism is semantically applicable in a sufficiently specified possible case. – from (3)

Argument from Division of Linguistic Labour

(1) If one is not an expert in a given area, then one does not know all semantic application conditions of specialist terms from that area.

(2) If one does not know all semantic application conditions of a term, then one is not able to tell whether the term is semantically applicable in at least some sufficiently specified possible cases.

(3) If one is not an expert in a given area, one is not able to tell whether a specialist term from that area is semantically applicable in at least some sufficiently specified possible cases. – from (1) and (2)

The way I interpret the argument from Reference Magnetism is based on the assumption that when scientists discover that the essence of the reference magnet associated with a term has certain specifics, having these specifics becomes a necessary and sufficient condition for the semantic application of the term. Thus, when scientists discovered that the microstructure of the liquid in our rivers and seas is H_2O, being a liquid composed of H_2O molecules became a necessary and sufficient condition for the semantic application of the term "water". Note that when having the newly discovered specifics of the essence of the reference magnet becomes a necessary and sufficient condition for the semantic application of the term, the semantic application conditions previously associated with the term are not rejected. Rather, they are *specified*. When scientists discovered that the microstructure of

[11] In using the terms "reference magnetism" and "division of linguistic labour", I adhere to the terminology employed by Haslanger (2012). Thus, reference magnetism as understood in the present study is not directly related to the view regarding the nature of reference that has its origins in the work of David Lewis and has become known under the same name. Rather, the argument from reference magnetism and the argument from division of linguistic labour correspond to two kinds of semantic externalism that are usually referred to as "environmental externalism" and "social externalism" respectively. It must be noted, however, that the argument from division of linguistic labour represents only a milder version of social externalism as opposed to the more radical version that consists in anti-individualism about meaning (discussed in section 2.2 of the present text).

water is H_2O, they did not deny that water is anything which is the same as the liquid in our rivers and seas. Rather, they specified what the liquid in our rivers and seas consists of. As far as the meaning of a term is the set of its semantic application conditions, we can say that such a specification of the semantic application conditions of a term constitutes a specification of its meaning. Thus, as far as meaning specification is a kind of meaning change, we can say that scientific discoveries of the said kind entail a change in the meaning of the given term.

Kripke (1980, pp. 119, 138) and Putnam (1975, p. 141) explicitly state that in their view, scientific discoveries of the said kind do not entail a change in the meaning of the given term, but they seem to be referring specifically to the case of meaning change in which the given term is not originally characterised by Reference Magnetism. Yet, the Kripkean view that statements about the essence of natural kinds like "water is H_2O" are necessary a posteriori seems to be inconsistent with the assumption that the meaning change described above occurs. After all, the meaning of a term is supposed to be known a priori. I think, however, that the supposed inconsistency is only apparent. Statements about the essence of natural kinds are originally empirical, i.e. a posteriori, but if the meaning change described above does occur, then they become a priori. It was an empirical discovery that water is H_2O, but if the meaning change described above does occur, then anyone who learns the meaning of the term "water" following this discovery learns a priori that water is H_2O. The assumption that the said meaning change occurs does not entail that such a statement is a posteriori and a priori at the same time – it entails that such a statement is first a posteriori and then a priori. In any case, I think that the question whether the said meaning change does occur is a question of the semantic application of the term "meaning change" (and possibly other related terms). In my own idiolect, this term is semantically applicable in the said kind of cases, and at least some other authors (e.g. Carnap, 1934; Glock, 2017) seem to agree.

If the said meaning change does occur, this has the following implications for the argument from Reference Magnetism. As I noted above, if it is true that when scientists discover that the essence of the reference magnet associated with a term has certain specifics, having these specifics becomes a necessary and sufficient condition for the semantic application of the term, then knowledge of the meaning of the term after the discovery entails knowledge of the specifics of the essence of the corresponding reference magnet. Thus, one does not know the specifics of the essence of the reference magnet, only if one does not know the meaning of the term. However, the experts in the given area do know the meaning of the term. Therefore, one does not know the specifics of the essence of the reference magnet, only if one is not an expert in the given area. Thus, the argument from Reference Magnetism can be interpreted or understood as a special case of the argument from Division of Linguistic Labour, or in other words, the argument from Reference Magnetism can be *reduced* to the argument from Division of Linguistic Labour.

In discussing the argument from Division of Linguistic Labour, we need to take into account the fact that not being able to tell whether a term is semantically applicable in a possible case makes conducting conceptual analysis of the term in a given language problematic, only if the term is a part of that language. Not being able to tell whether the term "beech" is semantically applicable to a given tree makes conducting idiolectal conceptual

analysis of the term problematic, only if the term is a part of one's idiolect. However, we can say that if one is not an expert in a given area (and thus does not know all semantic application conditions of specialist terms from that area), then the specialist terms from that area are not a part of one's idiolect. Experts determine the semantic application conditions of specialist terms in the communal language, and not in the idiolects of particular speakers. One might decide that in one's idiolect, the term "beech" would refer to one's dog, etc.[12] Thus, not being able to tell whether a specialist term from an area one is not an expert in is semantically applicable to a given possible case makes only conducting communal conceptual analysis of the term problematic, and not idiolectal conceptual analysis. Accordingly, we can say that the argument from Division of Linguistic Labour applies only to communal conceptual analysis and not to idiolectal conceptual analysis. And as far as the argument from Reference Magnetism is a special case of the argument from Division of Linguistic Labour, it applies only to communal conceptual analysis and not to idiolectal conceptual analysis as well. Thus, the whole criticism of conceptual analysis based on semantic externalism applies only to communal conceptual analysis and not to idiolectal conceptual analysis.

4.2. Experimental Philosophy

Experimental philosophy is a relatively recent approach to philosophy which aims to incorporate empirical studies into philosophical inquiry. A lot of the research that has been done within experimental philosophy is related to judgements about possible cases, which experimental philosophers usually call "intuitive judgements" or simply "intuitions". Sytsma and Livengood (2016, p. 36) make a distinction between "non-intuitional", and "intuitional" experimental philosophy and then divide the latter into two approaches – "the neutral approach", and the "evidential approach". The evidential approach is concerned with the evidential value of intuitions, and is itself divided into two programs – "the negative program", and "the positive program". The negative program is one of the areas within experimental philosophy that have received most scholarly attention (see Knobe & Nichols, 2017). It aims to demonstrate that judgements about possible cases are not reliable evidence and can thus be interpreted as directed against the method of cases in general and conceptual analysis in particular. Experimental philosophers have put forward two distinct arguments to the conclusion that judgements about possible cases are not reliable evidence (see Fisher, 2015; Knobe & Nichols, 2017; Horvath & Koch, 2021). They can be called "argument from diversity" and "argument from sensitivity" respectively and formulated as follows.

Argument from Diversity

(1) Judgements about possible cases vary between social groups.

(2) If judgements about possible cases vary between social groups, then they are not reliable evidence.

(3) Judgements about possible cases are not reliable evidence. – from (1) and (2)

[12] Very often, one would ascribe to them conditions that are only necessary for their semantic application, but not sufficient. Thus, a person who is not an expert in botany might ascribe to the term "beech" in their idiolect only the necessary semantic application condition of being a deciduous tree.

Argument from Sensitivity

(1) Judgements about possible cases are sensitive to irrelevant factors.

(2) If judgements about possible cases are sensitive to irrelevant factors, then they are not reliable evidence.

(3) Judgements about possible cases are not reliable evidence. – from (1) and (2)

Research in experimental philosophy has provided evidence in support of the first premise of each of these two arguments. The statement that judgements about possible cases vary between social groups is supported by studies which suggest that these judgements are influenced by factors such as cultural background (Weinberg et al., 2001; Machery et al., 2004) and gender (see, e.g. Buckwalter & Stich, 2014; Friesdorf et al., 2015), whereas the statement that judgements about possible cases are sensitive to irrelevant factors is supported by studies which suggest that these judgements are influenced by factors such as the emotions induced by a possible case (Cameron et al., 2013), the order in which possible cases are presented (Petrinovich & O'Neill, 1996; Swain et al., 2008; Wright, 2010) and the way an outcome is described (e.g. Petrinovich & O'Neill, 1996; Schwitzgebel & Cushman, 2015). There is some discussion about the extent to which the results from these studies can be generalized (Knobe & Nichols, 2017), but on the whole, we can say that the first premise of each of the two arguments is supported by the empirical data. The second premise of each argument, on the other hand, has been a subject of theoretical discussions.

The second premise of the argument from diversity is justified by arguing that if judgements about possible cases vary between social groups, then they are not reliable evidence, because one cannot assume that one's own judgements about possible cases are the same as those of people from different social groups (see Weinberg et al., 2001; Machery et al., 2004). If the method of cases is interpreted as conceptual analysis, then the same argument can be formulated as follows. If judgements about the semantic application of terms in possible cases vary between social groups, then they are not reliable evidence, because one cannot assume that one's own judgements about the semantic application of terms in possible cases are the same as those of people from different social groups, i.e. one cannot assume that the semantic application conditions one ascribes to a term are the same as those ascribed to the term by people from different social groups. However, one should make this assumption only if one is interested in the meaning of the given term in the communal language and not in one's idiolect. Thus, we can say that the argument from diversity applied only to communal conceptual analysis and not to idiolectal conceptual analysis.

The second premise of the argument from sensitivity is, in turn, justified by arguing that if judgements about possible cases are sensitive to irrelevant factors, then they are not reliable evidence, because one does not know whether a given judgement about a possible case is influenced by such irrelevant factors or not. If the method of cases is interpreted as conceptual analysis, then the same argument can be formulated as follows. If judgements about the semantic application of terms in possible cases are sensitive to irrelevant factors, then they are not reliable evidence, because one does not know whether a given judgement about the semantic application of a term in a possible case is influenced by such irrelevant factors as opposed to being based solely on one's procedural knowledge of the meaning of the term. It seems to me that this argument does apply to both communal and idiolectal conceptual

analysis. One's judgements about possible cases can be influenced by irrelevant factors, even if one is interested in the meaning of the given term in one's idiolect. However, if one is interested in the meaning of the term in one's idiolect, then one has the opportunity to easily put oneself in different situations and observe how these situations affect one's judgements. Thus, one can discover the effect of different irrelevant factors on one's judgements and possibly neutralise it. Thus, we can say that the argument from sensitivity undermines idiolectal conceptual analysis to a lesser extent than it does communal conceptual analysis.

5. Conclusion

Conceptual analysis is a method that consists in using judgements about the semantic application of a term in possible cases as evidence for a conclusion about the semantic application conditions of the term. The conceptual analysis in which one draws a conclusion about the semantic application conditions of a term in the communal language can be called "communal conceptual analysis", whereas the conceptual analysis in which one draws a conclusion about the semantic application conditions of a term in the idiolect of a particular speaker can be called "idiolectal conceptual analysis". In the present paper, I provided a systematic exposition and defence of idiolectal conceptual analysis, considering its possible and actual use in philosophical inquiry, as well as some of the most important criticisms that can be raised against it.

The two most influential kinds of criticism that have been raised against conceptual analysis in general are based on semantic externalism and empirical studies of judgements about possible cases respectively. Each of these two kinds of criticism can be understood as consisting of two separate arguments. The criticism based on semantic externalism can be understood as consisting of an argument according to which the semantic application of terms in possible cases depends not only on one's semantic competence, but also on empirical facts, and an argument according to which the semantic application conditions of some terms are known only to certain experts. I argued that the former argument can actually be reduced to the latter and that the latter argument applies only to communal conceptual analysis and not to idiolectal conceptual analysis. The criticism based on empirical studies of judgements about possible cases, on the other hand, consists of an argument according to which judgements about possible cases are unreliable evidence because they vary between social groups, and an argument according to which judgements about possible cases are unreliable evidence because they are sensitive to irrelevant factors. I argued that the former argument is applicable only to communal conceptual analysis and that the latter argument undermines idiolectal conceptual analysis to a lesser extent than it does communal conceptual analysis. Thus, we can say that on the whole, the two most influential kinds of criticism of conceptual analysis apply mainly to communal conceptual analysis and not to idiolectal conceptual analysis.

However, there is some evidence that most well-known and widely discussed instances of using conceptual analysis in philosophical inquiry have actually been instances of communal conceptual analysis. Communal conceptual analysis does have a certain advantage over idiolectal conceptual analysis, which consists in the fact that it can be used for resolving conceptual disagreements in the sense of demonstrating that one side in the disagreement is "right" and the other side is "wrong". However, idiolectal conceptual analysis can be used for

resolving conceptual disagreements in the sense of demonstrating that the two sides ascribe to terms different meanings. Furthermore, one can use idiolectal conceptual analysis in order to acquire propositional knowledge of the ordinary and technical meanings of terms in one's idiolect. Acquiring such knowledge can help one explicate one's own philosophical theories and arguments and thus understand them better, find potential faults in them and make them accessible to others. Thus, we can say that the results from the present inquiry suggest a shift of focus from communal conceptual analysis to idiolectal conceptual analysis. As far as idiolectal conceptual analysis has no practical use in philosophical inquiry for criticising the views of other philosophers and defending one's own views from criticism, we can say that this shift of focus would encourage cooperation and mutual understanding as opposed to trying to prove each other wrong.

References

Austin, J. L. (1966). Three Ways of Spilling Ink. *The Philosophical Review*, 75(4), 427–440.

Ayer, A. J. (1956). *The Problem of Knowledge*. Penguin Books Ltd.

BonJour, L. (1998). *In Defense of Pure Reason*. Cambridge: Cambridge University Press.

Braddon-Mitchell, D. & Nola, R. (2009). Introducing the Canberra Plan. In D. Braddon-Mitchel & R. Nola (Eds.), *Conceptual Analysis and Philosophical Naturalism* (pp. 1–20), MIT Press.

Buckwalter, W. & and Stich, S. (2014). Gender and Philosophical Intuition. In J. Knobe & S. Nichols (Eds.), *Experimental Philosophy Volume 2* (pp. 307–346), New York: Oxford University Press.

Burge, T. (1979). Individualism and the Mental. *Midwest Studies in Philosophy*, 4, 73–121.

Cameron, D., Payne, K. & Doris, J. (2013). Morality in High Definition: Emotion Differentiation Calibrates the Influence of Incidental Disgust on Moral Judgments. *Journal of Experimental Social Psychology*, 49(4), 719–725.

Carnap, R. (1934). *Logical Syntax of Language*. Kegan Paul, Trench, Trubner & Co Ltd.

Cekiera, K. (2023). Thought Experiments and Conceptual Analysis in Ethics. *Studia Philosophica Wratislaviensia*, 18(1), 29-43.

Chalmers, D. (2006). The Foundations of Two-Dimensional Semantics. In Garcia-Carpintero, M. & Macia, J. (Eds.), *Two-Dimensional Semantics* (55–140), Clarendon Press, Oxford.

Chalmers, D. (2011). Verbal Disputes. *Philosophical Review*, 20(4), 515–566.

Chalmers, D. & Jackson, F. (2001). Conceptual Analysis and Reductive Explanation. *The Philosophical Review*, 110(3), 315–360.

Chisholm, R. M. (1957). *Perceiving: a Philosophical Study*. Cornell University Press, Ithaca, New York.

Comesaña, J. (2005). Unsafe Knowledge. *Synthese*, 146(3), 395–404.

Davidson, D. (1987). Knowing One's Own Mind. *Proceedings of the American Philosophical Association*, 60, 441–458.

Descartes, R. (1641/1986). *Meditations on First Philosophy* (trans. J. Cottingham). Cambridge: Cambridge University Press.

Devitt, M. & Sterelny, K. (1999). *Language and Reality*. Blackwell Publishers Ltd, Oxford.

Farkas, K. (2008). *The Subject's Point of View*. Oxford: Oxford University Press.

Finlay, S. (2014). *Confusion of Tongues: A Theory of Normative Language*. Oxford University Press.

Fisher, J. (2015). Pragmatic Experimental Philosophy. *Philosophical Psychology*, 28(3), 412–433.

Fodor, J. (1981). *Representations: Philosophical Essays on the Foundations of Cognitive Science*. The Harvester Press, Brighton.

Friesdorf, R., Conway, P. & Gawronski, B. (2015). Gender Differences in Responses to Moral Dilemmas: A Process Dissociation Analysis. *Personality and Social Psychology Bulletin*, 41(5), 696–713.

Gettier, E. (1963). Is Justified True Belief Knowledge? *Analysis*, 23(6), 121–123.

Glasgow, J. (2008). On the Methodology of the Race Debate: Conceptual Analysis and Racial Discourse. *Philosophy and Phenomenological Research*, 76(2), 333–358.

Glock H. J. (2011). A Cognitivist Approach to Concepts. *Grazer Philosophische Studien*, 82, 131–163.

Glock, H. J. (2017). Impure Conceptual Analysis. In *The Cambridge Companion to Philosophical Methodology* (pp. 77–100), Cambridge University Press, Cambridge.

Goldman, A. (1976). Discrimination and Perceptual Knowledge. *The Journal of Philosophy*, 73(20), 771–791.

Grice, P. (1950/1989). Meaning. In *Studies in the Ways of Words* (pp. 213–223), Harvard University Press.

Grice, P. (1958/1989). Postwar Oxford Philosophy. In *Studies in the Ways of Words* (pp. 171–180), Harvard University Press.

Grice, P. (1989). *Studies in the Ways of Words*. Harvard University Press.

Grindrod, J. (2023). A Defence of Conceptual Analysis As a Linguistic Endeavour. *Theoria*, 89(4), 516–534.

Haslanger, S. (2012). *Resisting Reality: Social Construction and Social Critique*. Oxford University Press.

Henderson, D. & Horgan, T. (2011). *The Epistemological Spectrum: At the Interface of Cognitive Science and Conceptual Analysis*. Oxford University Press.

Horvath, Koch, (2021). Experimental Philosophy and the Method of Cases. *Philosophy Compass*, 16(1), e12716.

Horwich, P. (2020). Languages and Idiolects. In A Bianchi (Ed.), *Language and Reality from a Naturalistic Perspective: Themes from Michael Devitt* (pp. 285–297), Springer.

Ichikawa, J. J. & Steup, M. (2018). The Analysis of Knowledge. *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = ¡https://plato.stanford.edu/archives/sum2018/entries/knowledge-analysis/¿.

Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly*, 32, 127-136.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Clarendon Press.

Jackson, F. (2011). On Gettier Holdouts. *Mind and Language*, 26(4), 468–481.

Kipper, J. (2012). *A Two-Dimensionalist Guide to Conceptual Analysis*, Ontos, Frankfurt.

Knobe, J. & Nichols, S. (2017). Experimental Philosophy. *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), URL = ¡https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/¿.

Kölbel, M. (2021) Varieties of Conceptual Analysis. *Analytic Philosophy*, 64(1), 20–38.

Kornblith, H. (1998). The Role of Intuition in Philosophical Inquiry. An Account with No Unnatural Ingredients. In M. DePaul & W. Ramsey (Eds.), *Rethinking Intuition* (pp. 129–141), Lanham, MD: Rowman & Littlefield.

Kriegel, U. (2017). Metaphysics and Conceptual Analysis: Experimental Philosophy's Place under the Sun. In D. Rose (Ed.), *Experimental Metaphysics* (pp. 7–46), New York: Bloomsbury Academic.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Laurence, S. & Margolis, E. (2003). Concepts and Conceptual Analysis. *Philosophy and Phenomenological Research*, 67(2), 253–282.

Lewis, D. (1994). Reduction of Mind. In S. Guttenplan (Ed.) *Companion to the Philosophy of Mind* (pp. 412–431), Blackwell.

Machery, E., Mallon, R., Nichols, S. & Stich, S. (2004). Semantics, Cross-cultural Style. *Cognition*, 92(3), 1–12.

Moore, G. E. (1903). *Principia Ethica*, Cambridge: Cambridge University Press.

Nolan, D. (2009). Platitudes and Metaphysics. In D. Braddon-Mitchell & R. Nola (Eds.), *Conceptual Analysis and Philosophical Naturalism* (pp. 267–300), The MIT Press, Cambridge.

Overton, J. (2013). 'Explain' in Scientific Discourse. *Synthese*, 190(8), 1383–1405.

Petrinovich, L. & O'Neill, P. (1996). Influence of Wording and Framing Effects on Moral Intuitions. *Ethology and Sociobiology*, 17(3), 145–171.

Plato. Theaetetus (translated by B. Jowett). In B. Jowett (Ed.), *The Dialogues of Plato*, Oxford: Oxford University Press, 1892.

Putnam, H. (1973). Meaning and Reference. *The Journal of Philosophy*, 70, 699–711.

Putnam, H. (1975). The Meaning of 'Meaning.' In *Mind, Language and Reality: Philosophical Papers Vol. 2*, (pp. 215-271). Cambridge University Press.

Quine, W. O. (1951). Two Dogmas of Empiricism. *Philosophical Review*, 60(1), 20–43.

Quine, W. O. (1960). *Word and Object*, The MIT Press, Massachusetts.

Ryle, G. (1949/2009). *The Concept of Mind*, Routledge.

Schroeter, L. (2021). Two-Dimensional Semantics. *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.), URL = ¡https://plato.stanford.edu/archives/fall2021/entries/two-dimensional-semantics/¿.

Schwitzgebel, E, and Cushman, F. (2012). Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-philosophers. *Mind & Language*, 27(2), 135–153.

Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3, 417–57.

Sekowski, K. (2022). Concept Revision, Concept Application and the Role of Intuitions in Gettier Cases. *Episteme*, 21(3), 901–919.

Strawson, P. (1950). On Referring. *Mind*, 59 (235), 320–344.

Strevens, M. (2019). *Thinking Off Your Feet: How Empirical Psychology Vindicates Armchair Philosophy*, Cambridge, Massachusetts: Harvard University Press.

Stuart, M. (2015). Philosophical Conceptual Analysis as an Experimental Method, In T. Gamerschlag, D. Gerland, R. Osswald & W. Petersen (Eds.), *Meaning, Frames, and Conceptual Representation* (pp. 267–292), Düsseldorf University Press, Düsseldorf.

Sytsma, J. (2010). The Proper Province of Philosophy: Conceptual Analysis and Empirical Investigation. *Review of Philosophy and Psychology*, 1, 427–445.

Sytsma, J., Livengood, J. (2016). *The Theory and Practice of Experimental Philosophy*, Broadview Press, Peterborough.

Tallant, J., Andow, J. (2020). English Language and Philosophy. In S. Adolphs & D. Knight (Eds.), *The Routledge Handbook of English Language and Digital Humanities*.

Valchev, H. (2018). What Is Conceptual Analysis? *Balkan Journal of Philosophy*, 10(2), 131–142.

Valchev, H. (2022). Empirical Conceptual Analysis: An Exposition. *Philosophia*, 50, 757–776.

Valchev, H. (2023). Does Scientific Conceptual Analysis Provide Better Justification than Armchair Conceptual Analysis? In D. Bordonaba-Plou, (Ed.), *Experimental Philosophy of Language: Perspectives, Methods, and Prospects* (pp. 57–74), Springer.

Weinberg, J., Nichols, S. & Stich, S. (2001). Normativity and Epistemic Intuitions. In J. Knobe & S. Nichols (Eds.), *Experimental Philosophy* (pp. 17–46), Oxford University Press.

Wikforss, Å. (2017). Incomplete understanding of concepts. *Oxford Handbooks Online*. https://doi.org/10.1093/oxfordhb/9780199935314.013.49.

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford University Press, New York.

Williamson, T. (2007). *The Philosophy of Philosophy*. Malden, MA: Blackwell.

Wittgenstein, L. (1951). *Philosophical investigations*. Oxford: Blackwell.

Wright, J. (2010). On Intuitional Stability: The Clear, the Strong, and the Paradigmatic. *Cognition*, 115(3), 491–503.

Zach, M. (2019). Conceptual Analysis in Philosophy of Science. *Balkan Journal of Philosophy*, 11(2), 107–124.

Balkan Journal of Philosophy (2025), vol. 17, issue 2, 186–198

CAN THE SEARLEAN NOTION OF COLLECTIVE INTENTIONALITY AND POWER LEAD TO COERCION?

Neetika Singh*

Lalit Saraswat**

Vineet Sahu***

Abstract: John Searle argues for a notion of 'collective intentionality' that produces social actions with the attributed deontic status function on an external reason, simultaneously recognized by all the members. Social actions depend on the acceptance of a propositional attitude in the light of a constitutive rule and the 'we intention' of each member. However, Searle's intentional theory of actions argues that human beings are rational in nature and freely choose to act as per a desire independent of reason; hence, offering a possibility of unreflective obedience towards social rules. What implications and considerations may emerge while explicating relationships between intentionality, intentional action, the notion of power, and coerced life? This paper assesses the scope of such collective participation (in the Searlean sense) that boils down to a coerced life. The viewpoint of a coerced life is examined along with the notion of intentionality and background, and how these notions translate into a concept of power.

Keywords: Collective Intentionality, Power, Background, Coercion, Agency.

1. Introduction

Aristotle's *koinonia*, Rousseau's *volonté générale*, and/or *people's* spirit are terms that approach what is labeled as 'collective intentionality.' Despite various implied versions and viewpoints, a proper comprehensively coherent account of 'collective intentionality' has not remained unassailable. Also, there have been some other differing and contrasting ways to deal with this notion, like Durkheim's (1898) and Weber's (1922). Durkheim mainly focuses on consciousness and emotional aspects, rather than attitude, and the notion of the collective is 'controlling' in nature.

A proper way to explicate the human social ontological structure is to account for the analyses of some properties of the mind that create such an ontological realm. Searle's assertion on the ontological foundation of the social world is grounded in a rule-based (regulative and constitutive) structural analysis with intentional actions. Searle (2010) holds that our account of the world should focus on how 'we live in exactly one world' rather than the postulated Worlds. It is not an ontologizing of the monistic viewpoint nor the rejection of dualism, trialism, etc. for that sake. The only thing this account suggests is that we regard

*Department of Humanities and Social Sciences, IIT Kanpur

**Department of Humanities and Social Sciences, IIT Kanpur; lalits@iitk.ac.in

***Department of Humanities and Social Sciences, IIT Kanpur; vineet@iitk.ac.in

the basic structures and facts endorsed by physical, biological (evolutionary), and other natural sciences. So, from the subatomic levels to societies and proteins to prime ministers, all accounts of our investigations must agree with the most basic requirement: 'the consistency with the basic facts and demonstration of how the non-basic facts are dependent on and derived from the basic facts' (Searle, 2010, p. 4). Searle argues for an object of knowledge that is ascribed to the deontic status function and recognized as the external reason by all members. Social action depends on the acceptance of a *proposition* in the light of a constitutive rule and the 'we intention' of each member. Group members associate with each other based on a dispositional belief about each other's we-intentions. His intentional theory of actions argues for human beings as rational beings who freely choose, yet pertain to the possibility of unreflective obedience towards social rules. The scope of such collective participation leads to a coerced life, and the only way to deal with such deviants is to use power against them. Such a method may be justified because there are people appointed by/as authorities to follow this dutiful act. But, it is plausible that the Searlean account thus boils down to an account of a coerced life. To elucidate this, the concept of 'collective intentionality', background, and the notion of power are examined from Searle's perspective.

Regarding the meaning and conceptual significance of the term "lead", in connection with the concept of intentionality, power, and coerced life, it is not a logical entailment for coercion; rather, there exists a possibility of a coerced life for the members of a collective under the collective intentionality-driven status functions, rules, etc., and it is therefore essential to understand what justifies coercion, or how coercion seeped into and is justified according to a Searlean perspective. The argument is based on Searle's notion of bad faith, which exists among members of a community due to the possibility of an unreflective, yet rational behaviour opted by them, which is caused by the structured background of consciousness. This raises the question: is organized coercion necessary in order to sustain certain systems of status function? In his theory, the dilemma of acting in the social world is bracketed only under obligatory social actions. Searle's collective intentionality with active and inactive agents cannot afford to have a dilemma beyond deontological concerns, if it is possible at all. The rational or active agents identify the obligation and act reasonably. Passive agents, though, at one end perform their duty by recognizing external desire-independent reasons, yet are ignorant of the purpose of their action. The latter have been trained mechanically to identify the connection between external reason and obligation. The agents who do not follow this pattern should be ready for forced attitudes from authorities. Skills are practiced in such a way that dispositions are acquired for responding appropriately, as determined by the rules. The authority to decide on the appropriateness of action lies with the structure rather than the members. In this paper, we also examine whether and in what manner, in an institutionalized world, power gets justificatorily generated from external control(s).

2. The Constitution of 'Collective Intentionality'

Searle argues that intentions are *sui generis* as they are verified based on the corresponding physical movement. All intentions are based on an external reason to act, and a 'we-intention' is a type of intention. Searle narrates two scenarios: first, where a few people are sitting on the grass in different places in a park and they start running simultaneously towards a shelter. Each individual has an intention of 'I am running towards the shelter'. And the second scenario where a group of people is part of a ballet and the choreography

calls to converge at a certain point. If observed by an outsider, both the behaviors would appear collective (as the physical movements of people appear to be directed towards a common goal). However, it is not so, because these actions have two components viz. internal mental states, i.e., intention plus external bodily movement.

If explained, the intention of individuals can be modified to an 'I intention', for example – each individual recognizes rain as an external reason, develops a proposition "I am running towards a shelter", and thereby performs a physical movement per it. Every individual's intention is concerned with the objective external reason and not something that is specifically in relation to others, even though they may be aware of others' intentions as well; whereas, in the second case, each individual has a mental state of a 'we intention' to perform the ballet together. This 'we intention' is not based on an individuated reason to dance but to dance as part of a group. It is important to note here that per Searle, every individual performing a ballet in the group: (1) recognizes the command of choreography with a particular imposed status function with a rule: 'X counts as Y in context C' (in this case, dancing counts as a ballet performance in the context of a group) and (2) participants have a 'we intention' and are acting as a part of the group. Each individual converts this external reason into the proposition: "we are dancing in a ballet" and performs it physically.

However, this example requires further analysis. Searle means that an internal mental state can cause collective behaviour. The similarity between the two cases is that the bearers of intentions are individuals. The 'we intention' of performing in a ballet exists separately within different participants. Each member has his/her own 'we intention'. The ontology of the 'we intention' is that it exists within the mental state of the individual and not the collective. No single mind is encompassing all minds as one collective mind. There are plural minds with their mental states. An 'I intention' is borne by a person; similarly, a 'we intention' can be claimed by only an individual. Searle's notion of 'we intention' can be understood in terms of propositional mental states and physical movements that are limited to individuals only. This is because: the external reason is solely recognized and translated into the proposition by the individual. Others do not share it as each has its own ontologically separate mind to 'we intend'. If a person says that she 'we intend' to dance in the ballet, then it is only she who can claim direct access to that 'we intention'. "All intentionality, whether collective or individual, has to exist inside individuals' heads" (Searle, 2010, p. 44). It can be said that because an individual's direct mental access to their intentions corresponds to a first-person ontology, Searle's theory of intentions as a personal mental component of an individual's action is closely related to the theory of Ontological Individualism. It is the mental state and behavior of individuals that can potentially explain the construction of the social world. However, this is not the only factor necessary for understanding the social world, and thus his theory leaps ahead of ontological individualism.

Searle argues that in a group, each person acts with a 'we-intention' in the mind and the awareness that others have their 'we-intention'. There are several minds with each possessing direct access to their minds. And with the actions of other participants or observables, they can infer that others are also having a 'we-intention' to act in favour of a rule. They all recognize and choose the same external reason to be converted into the proposition of a 'we intention' of all members. They are all *Collectively Rational* individuals in terms of one external reason recognized by those involved in a group so that they act in a direction of

fit, i.e. mind-to-world. Or, it can be said that all the members recognize the same external reason, i.e. the commands of the choreography to act in its favor. They all (1) believe that the fellow members accept the authority of choreography by recognizing it as the external desire-independent reason (choreography and its commands); (2) convert this belief into a proposition of a 'we intention' to act.

Another factor that Searle adds to explain the ontological status of the collective is the need for a certain type of psychological disposition. Searle's social individual with a 'we intention' believes that others are thinking with a 'we intention' and are committed to acting according to the norms of a group (Searle 2010, p. 45). Collective action is a normative behavior of all. However, each individual only has command over his/her mental state and not others'. At most, they can assume and believe that since others are following norms, then they must be acting with a 'we intention'. Chronologically, a belief in others' involvement is assumed first and then later members can infer others' involvement from their actions.

Suppose an individual follows all the norms of the group – would that mean that s/he is acting with a 'we intention' only? What if the person is following them with an intention of securing exclusive benefits for oneself rather than others? Searle would say that, at least, they believe in being aware that they are a part of a group. To attribute to an action the status of a collective action, all members must have a 'we intention', which is about the external reason. The belief of the invariable[1] external reason and conversion into the proposition of a 'we intention' forms a connection between individuals of a group. The nature of the 'we intention' is that it is inclusive of others' recognition of the objective desire-independent reason. The objectivity of external reason makes the 'we intention' collective. Although the propositional content of the 'we intention' is about objective desire-independent reason, the ability to form such a proposition is limited to the individual ontologically holding it. Suppose there are two individuals in a group then:

Collective Intentionality

'We intention' of individual 1 + 'We intention' of individual 2

(Members are directed together at)

A mutual belief that both refer to the external desire independent reason

Searle states that "... Collective mental phenomena of the sort we get in organized societies are themselves dependent on and derived from the mental phenomena of individuals" (Searle, 2010, p. 4). Whether it is an 'I intention' or 'we intention', both are mental phenomena that belong to a first-person ontology. And so, because there is always a possibility of ignorance of the content of others' intentions, Searle's theory claims subjective ownership of an individual's content of intention or *Subjective Individualism* (Hans Bernhard Schmid,

[1]We are using the term 'invariable' to clarify that all the members perceive an object in the social world to be identified as an external reason and oblige to its status function. The object does not vary or change for different members. Suppose a stone is taken to be a border in a collective – then all members conceive it as a border. Similar is the case with money and traffic lights, etc. Money is a functional concept attributed to paper notes. People may hold different notes, but all the notes stand for the institutionally-imposed social concept of money.

2003). For Searle, the mental phenomenon, ontologically, does not belong to the group but to each member. The number of 'we intentions' is the same as the number of separate mental phenomena of respective group members. Searle's theory keeps both the 'we intention' and 'I intention' in the same category of mental phenomena, i.e., they bring local or personal effects through one's physical movement. Collective action is the sum of the 'we intention' and the corresponding local or physical movements of each member. The direction of fit of intention in action is world-to-mind, and the direction of causation is of mind-to-world. The intention-in-action is the cause of the agent's movement. The intention-in-action causes a local or personal effect of one's bodily movement (Searle, 1983, p. 95) and not the overall goals or purposes that such movements are supposed to bring about. A narrow sense of social action is explained by Searle's theory of intentionality:

> ...The propositional content can only represent the condition of satisfaction of the intention. The general point is that the propositional content of an intentional state always specifies the condition of satisfaction as distinct from the type of state that it is, which is specified outside the propositional content ...the propositional content can only represent elements that an agent can (or thinks he can) causally influence. (Searle, 2010, p. 45).

Propositional content within the 'we intention' is available to the other through its condition of satisfaction of the intention or the mental state i.e., the physical movement. Since one's intention is the cause of his/her action, others' intentions cannot be the cause of his/her actions. Individuals do not control each other's 'we intention' and its propositional content. Actions of each member of a group are an effect of separate 'we intentions.'

The connection between two individuals is the belief about possession of a 'we intention' with the assumption that both are referring to the same external reason inferred through each other's actions. There is a disposition of believing-that-p. Psychologically, all members hold a belief regarding the involvement of other 'we intention'-holding individuals involved in the group act that they are directed towards the same external factor, or desire-independent reason. Such dispositional capability of a participant (possessing a 'we intention') presupposes collectivity by taking others' 'we-intended' actions for granted.

It can be charged that the notion of 'collective intentionality' in Searle's theory commits to the fallacy of circularity. Searle would argue that for an individual to be a part of a group does not require that they hold a 'we intention'. But, because every human by natural construction of consciousness can have a 'we intention', therefore, people choose rationally (enabled by linguistic capability) to participate in group activities and their rules. The foundation of the 'we intention' is in the neurobiological capability of having 'we consciousness', enabled by language. Humans form a group because of this natural construction. For example, as Searle states, acknowledging the sports team and playing for it with a 'we intention' can be the reason for the cooperative behavior of team members. It is the commitment of all members to the external desire-independent factor. However, Searle's argument of collective intentionality seems inconsistent where it is stated that a participant (possessing a 'we intention') takes others' 'we intended' actions for granted in a group. Collectivity is a concept for Searle, in which each member is necessarily presupposing the others' involvement. To act with a 'we intention' according to group norms, each member assumes the concept of being together in it. The neurobiological emergence of 'we consciousness' is located in such

a presupposition of the agent. Agents require a leap of faith in others' involvement to engage in the group activity. This leap of faith can fall into bad faith sometimes (but the issue of bad faith will be discussed later in the section on rationality in social backgrounds).

Once the circularity notion has been discussed, we need to go back to Searle's notion of collective and shared visual experience.

3. The Rationality of Background Knowledge

John Searle's notion of shared experience in a collective is explained via shared visual experience. When people report visual experience, they are aware of a proposition along with an intention in action. Visual intentional content is self-referential, and the direction of causation is world-to-mind, and the direction of fit is mind-to-world. "... [W]hen I have a representation of an Intentional object in a belief or a desire it will always be represented under some aspect of other ... the aspect of visual perception is fixed by a sheer physical feature of the situation" (Searle, 1983, pp. 50–51). Searle argues that an individual always sees an object in a 'particular' aspect of the object of knowledge. Each person has direct access to the intentional content of oneself but not to that of others (even if it is a 'we intention'). Searle states that when two people have visual experiences of the same object, they need not require seeing it under the same aspect of the object because the act of seeing involves the shared visual experience of a particular object and not its aspects. Two people have shared visual experience as a part of 'our seeing it' i.e., a 'we intention'. But what makes one aware of the fact that the other is also seeing under the same aspect i.e., what about the commonality of the aspects concerns shared visual experience?

Searle states that "there can be a variety of different types of shared experiences, and am not sure how, or even if, the various complexities can be represented in the notation we have been using so far ..." (Searle 1983, p. 71). There is always a possibility of ignorance of various aspects of the object of knowledge. By aspects of the object of knowledge, Searle indicates here, not just the physical angle of the object itself but what meaning, or the propositional content of intentions in the minds of individuals, it evokes. Various aspects of an object of perception may have different meanings, but Searle is not concerned about that. However, if Wittgenstein's (2001) tricky picture of the rabbit and the duck is considered, it can be said that the same object presents two different visual experiences. One person can view one aspect at a time. Suppose every object of perception has different aspects and only one aspect can be seen at a time. In that case, people looking at it together can refer to the same object of perception but recognize its different aspects as a reason. The reference point is the same, but the meanings interpreted in their propositional content can be different for different people looking at it at the same time. Different individuals having indirect access to others' intentions (even if they are in the act of looking together at the same referred object) opens a possibility of ignorance of respectively different observed aspects of the object by each individual. Suppose a painting is looked at by two people as a team with 'we intention', then it need not be seen in the same aspect by viewers. In many cases, it may be true that perception corresponds to a propositional perception of the type 'we intend'-seeing-that-p. But if ignorance of aspects is added, then propositional content can be of a type, 'we-intend'-seeing-that-p-in-particular-aspect. The commonality of aspects as a factor in having a common goal needs to be ascertained in that case.

However, Searle (2010) states that there can be collective behavior where all members of the group contribute a 'we intention' that has different contents for different individuals, as in the case of a duet:

> In collective intentionality, it cannot be required of each individual's intentionality that he know what the intentionality on the part of others is. In complex forms of teamwork or collective behaviour, one does not know what others are doing in detail. All one needs to believe is that they share one's collective goal and intend to do their part in achieving the goal (Searle, 2010, p. 45).

The aspects of a common external reason they are looking at may be different, but they believe that they share a goal of collectivity. Belief in one's contribution towards collective performance through a 'we intention' is a sufficient criterion for a group activity. Each participant's propositional content may differ from others, as in a duet, because each participant is viewing the same musical set of notes yet playing different aspects of it. The third-person perspective '*they*' depends on the psychological belief in everyone's contribution to the collective. The actions are directed towards collaborative performance through contribution not to a further common goal of the group, but through their immediate and significant performance with a 'we intention'. A group forms when a member assumes a psychological connection with the other members of the group. There is a group psychology of coordination, where each is driven by one's collective orientation. The collective orientation is reflected in the immediate action of each member and not any goal beyond it. People cooperate because they recognize a socially-imposed function of an external factor that causally serves a purpose. Human beings are rationally collective beings because they can be directed towards the external reason together, intentionally impose a function and accept its obligatory nature.

Per Searle, human beings are thinking beings. We are rational as we intend to act in a certain way based on a chosen reason among several. The argument goes:

> Since, all humans are rational beings,
>
> And if suppose,
>
> Mila is a human, therefore,
>
> Mila is a rational being.

If there are actions, then there must be reasons chosen by the agent. The process of action includes recognition of an 'external reason' or 'desire-independent reason', such as 'wind blows in from the window', by an agent that is transformed into propositions (as the content of intentions). It is also necessary that each individual has such a mental state to act in a particular way. Searle puts forward the concept of 'intention-in-action' while analyzing the term 'action' (Searle 1983, pp. 83-98). An action is "a causal and Intentional transaction between mind and the world" (Searle 1983, p. 88); consisting of (1) an internal mental state called intention, which is appropriated with (2) an external physical movement. Intentions are the mental component that is accompanied by the movement (Searle, 2010, p. 51) Action is based on a reason with practice as a condition of satisfaction.

4. On the notion of Background

Searle states in his notion of background that a person "... develops skills and abilities that are ... functionally equivalent to the system of rules, without actually containing any representation or internalization of those rules" (Searle, 1995, p. 142). Skills are imbibed or practiced in such a way that dispositions are acquired to respond appropriately, where the structure of rules determines appropriateness. Searle argues that though rules are not as such internalized, nonetheless with practice they can be mechanically driven through structural configuration, where further active engagement with rules stops.

Searle's suggestion regarding a separate set of abilities that are functionally equivalent to rules acquired through practice makes room for inactive or passive agents. For Searle, there are parallel functional structures of 'Background and Intentionality'. It turns out after a point, rules are just followed, i.e., people simply "know what to do ... We do not apply rules consciously or unconsciously ..." (Searle, 1995, p. 143). So it becomes passive participation in the system of rules, which implies the mechanical approach towards actions. Searle could have said that we do not remind ourselves repeatedly of the rules and their purpose, even though we are aware of them. But he does not do that. He rejects the possibility of both conscious and unconscious engagement with practice once the background is set. And, therefore, he refers to the background as the realm of development of "mechanisms" that "has evolved precisely so that it will be sensitive to rules" (Searle, 1995, p. 146). These mechanisms aid intention to conform to the rule but not to aid the human experience of engagement with constant conscious effort or practice. At the functional level, humans develop a phenotype in response to the environment and learn to cope with it. The causal role of teleology to the rule is inverted and eliminated after humans absorb it. In a duet, participants are just governed by the skillful mechanism of playing instruments. They are trained in it, and to be a part of a duet, they need to bear a disposition that others are involved and nothing more. They oblige and display their expected skills. If there is sufficient causal reason then a member acts automatically, just like an addict. Habits are these kinds of compulsory actions.

The background is detachable from an ongoing process of action. A person behaves the way she does because she is following the social rules. This is to say: 1) the person behaves the way he does because she has a structure that disposes her to behave that way, and 2) she has come to be disposed to behave that way because that's the way that conforms to the rules of the institution. In other words, the agent loses awareness about the rules' purpose and does not need to know the teleological source or even reason for the formation of the rules of the institutions to follow them or to conform to the rules. However, the question of free rational choice arises here, as a subject has absorbed the rules without any reflection on their purposes. Acting, per will, can be applied with obligatory behavior only, i.e., the background is instilled in the individual in such a way that she must surrender to the rules and get governed by them. If one does not understand their purpose, they can and must follow them as almost a reflex or an automated responsive attitude without any conscious or unconscious effort. Searle admits that the issue of intelligibility of action is lost here (when constant engagement with rules is scrapped). It remains intelligible only in terms of binding. The history of events that brought the social rules into effect is detachable in Searle's theory of collective intentionality. If a person develops bad faith in a rule, Searle's theory does not

provide any way out so that s/he can overcome it herself/himself. There are authorities appointed to forcefully evict any disobedient agents.

5. Deontic Power and Passive Agency

In the process of intention formation and action, the first-person view is that actions are something that are done rather than events happening passively. A person cannot wait and see something happen, as in the case of a billiards game (Searle, 2001, p. 232). As discussed earlier, in Searle's concept of collective intentionality, each participant causes their action rather than others'. One does not wait for others to act first and to form the intention following their actions. Rather every individual assumes coordinated behavior in favor of a single external factor as a reason. Actions of others are not evaluated and based on external reasons to act; a 'we intention' is about the objective observed factor. The 'we intention' is directed towards one object for all group members rather than each other's intentions and actions or one's desire-dependent reasons. The collective behavior of the participants of a group is expected based on the recognition of invariably one external desire-independent reason (the object may have different aspects, but this is not a concern for Searle). The directedness of an intentional state is towards the same object in the case of all group participants rather than being varied. It would be appropriate to say that such a notion of 'collective intentionality' is based on a written or unwritten mutual agreement regarding the recognition of external reason and fulfillment of expected obligations[2]. Searle's theory of rule is about expectations. The expectation is that rules are to be obeyed and their status function is to be fulfilled. A function is about the particularity of roles or behaviors in accordance with the rules.

For Searle, reasons can be the ground of desires rather than vice versa (Searle, 2010, p. 129). Obligations regarding the social rule with status function imposition on an object of knowledge can be followed by a desire to follow it. Desires are subservient to the desire-independent reason or obligation. The constitutive rule of 'X counts as Y in context C' is obligatory because, in the social background, such a status function has been a collectively accepted and recognized fact. In the social world, the rule is where something called "X has been collectively accepted as Y because it happens to be in the context of the social world".

The term 'recognition' is crucial. It means that there is an object in the world that is represented as existing in the social context. The social object has been declared to have the functionality to an effect. The status functionality holds an obligation towards it because of the collective agreement on it. "When the representations are collectively recognized or accepted, the line of stones acquires a new status: it is now a boundary. And such representations exemplify precisely the logical form of the Status Function Declaration" (Searle, 2010, p. 96). These representations bear an authoritative orientation or declarative claim. They are binding on the agents of the group. There is a generalized law regarding the

[2]He explains in the example of Business School case 2 that an agreement among collective members is a necessary clause for their involvement in collective behavior. There occurs a *solemn pact* among graduates that they will serve humanity. Searle believes that such a pact or agreement, whether written or unwritten, is the ground of the social world. Members of a group are obligated to fulfill such a promise and since all believe or have good faith that others are involved volitionally, they tend to keep their promise. Such promises are either true or false – true if they present a condition of satisfaction, otherwise false. There is no possibility of a dilemma in such cases.

socially-imposed function on an object, and the ensuing obligatory status causally produces a desire to act per it. People rationally choose to act out of this desire for an independent reason if they recognize it.

But what if one does not recognize it, and hence does not follow such a rule? For Searle, there are two basic options regarding types of action, i.e. either to follow a rule or not. The first one has been discussed earlier, where each member of the group intentionally recognizes an object of knowledge as an external desire-independent reason and accepts its obligatory demands. In a second way: only the one who does not identify the obligatory or deontological status in the social world would not follow the rules. However, there is also a third option, that is, to unreflectively identify the social deontic status function and somehow act in its favor, because Searle specifically clarifies that a desire does not 'follow' from obligation, although it 'can'. There is a possibility of 'ought' to 'can', but the latter does not necessarily follow from the former. He points out a psychological 'gap', i.e. there occurs a lack of sufficient causal conditions to decide on something and act thereon. He explains that such a gap occurs but does not dwell on the question of what creates this gap.

With the possibility of a third option, a form of passive participation in a collective and its constituent rules is an option for Searle. Collective intentionality is composed of two types of agents viz.: 1) actively participating 2) passively participating:

> [M]any people simply go along, unreflectively, with social situations in which they find themselves. But this can amount to a form of inauthenticity or even bad faith, because they are creating desire independent reasons which are rationally binding on them but which they have created if they might have not thought about the question (Searle, 2010, p. 132).

There may be unreflective agents who simply abide by the social rules. Although they do not know why or on the basis of which fact, they are acting in such a way, yet these passively participating agents are still held responsible for being part of such a social system because they are the ones who are imposing these bindings on themselves. However, the unconscious or unreflective following of rules makes it a mechanical process. A passive agent creates the external factor as a desire (desire-independent reason to act), which may not be valid or verifiable. Let us remind ourselves of the basic criterion that the status function is collectively recognized and accepted as obligatory in the social milieu. Both types of rational agents follow that criterion. The only difference is that the latter are unreflective of it.

The two basic criteria of collective rational action are fulfilled by them, even though they do not focus on the question of the fact for which they are abiding by the rules. They create an option to act without knowing the source of it. The issue is: why do people adopt social rules before they have reflected on them, and why do they follow them when, at this stage, they are an unconscious or meaningless pursuit? They show the same behavior as a rational agent while contributing to collective intentionality. Their intentional state is accompanied by a propositional attitude similar to that of a rational agent. Even with the aspect of an external factor having a first-person ontology, the person having such unreflective propositional content possesses a 'we intention' to physically participate in a group activity. Others can infer coordination from the action supported by the self-referential intentional state, with the condition of satisfaction. To understand why Searle's theory of

collective intentionality allows for an option of unreflective collective participation, we have to understand the concept of background.

6. Coercion and Power

Searle justifies organized coercion as an acceptable way to curb disobedience. It is the only way left. He states, "... Organized coercion is indeed necessary to sustain certain systems of status function, but police forces, armies, and other forms of organized coercion are themselves systems of status functions" (Searle 2010, p. 142). Those who choose the status function, even unreflectively, and can recognize the external desire as an independent reason for action are rational. However, if the collective behavior of individuals is unreflective, then it devolves into an incoherent behavior. Searle's theory does not answer why people act unreflectively. Coherence or incoherence is not a criterion for Searle to call an action rational or irrational. He agrees that such behaviors become meaningless or unintelligible to the extent that people are merely being forced to follow rules.

The people who are meant to coerce irrational agents receive justification for their actions due to their deontic position. They are limited to the responsibility of generating their intention in action towards the rule and an effect on wrongdoers in that direction, nothing else. The force in their course of action is mandatory to disallow any breach of rules. Searle believes "only to the free agent do such systems make any sense, but precisely for free agents such systems are necessary" (Searle 2010, p. 142). He does not discuss any other way than coercively including these irrational agents in the collective. Punishments and rewards are the means through which involuntary and forced social actions can be introduced.

In Searle's theory, the dilemma of acting in the social world is bracketed only under obligatory social actions. Searle's collective intentionality with active and inactive agents cannot afford to have a dilemma beyond deontological concerns if it is possible at all. The rational or active agents identify the obligation and act reasonably. Passive agents, though at one end perform their duty by recognizing external desire and independent reasons, yet are ignorant of the purpose of their action. The latter have been trained mechanically to identify the connection between external reason and obligation. The agents who do not follow this pattern should be ready for forced attitudes from authorities. Skills are practiced in such a way that dispositions are acquired to respond appropriately where "appropriateness is actually determined by the structure of rules" (Searle, 1995, p. 142). The authority to decide on the appropriateness of action lies with the structure rather than the members. In an institutionalized world, power is justifiably generated from external control.

A critical assessment of the relationship between collective intentionality, domination, rules, norms, and the controlling aspects associated with power shows how it may define success. Similar to Thompson's (2017, p. 211) view of domination as a kind of power enabler amongst agents, Searlean deontic power not only has a scope of institutional facts, but Searlean *rules* may assign power, duties, and obligations to some members of the group:

> We (or I) make it the case by declaration that a Y status function exists in C and in so doing we (or I) create a relation R between Y and a certain person

> or persons, S, such that in virtue of SRY, S has the power to perform acts (of type) A (Searle, 2010, pp. 101–102).

The members of the group therefore start thinking, acting, and cognizing things, being rule-bound. The Searlean notion of social power is mostly Weberian (limited to agents, power over each other), but mainly cognitive and mental. In the Searlean account, coercion is possible when agents start internalizing the constitutive rules, minimizing the other-oriented actions that may benefit them (others).

7. Concluding discussion

The notion of collective rational action in Searle's theory of 'collective intentionality' is a narrow concept, and hence the theory of actions is restricted within the parameters of rule-following. Mechanisms of rule can only explain obedience and competitiveness, creating a possibility of passive cooperative activities. The concept of a constant, consciously engaging background of practices is eliminated, resulting in a coerced life for many. Due to the mechanistic nature of the background and the concept of the 'we intention' causing personal or local effect, agents or collective members do not have a way out of passive participation in the structure. It is the external force that can bring them in line with the obligatory behavior. However, due to the first-person ontological nature of the subject's 'we intention', authorities cannot make them understand the reason for their neglect, and thus the agents remain in an unreflective state. At most, authorities can justifiably intimidate them with punishments and even rewards. Searle's theory of the collective argues for a wider range of criteria for rational agents who are mechanically habituated to the norms of structure, within which there is a subclass of agents who are habituated and unreflective at the same time. In such a system, punishments are mandatory, and thus, a coerced way of life may find a way to become a norm of the social world. Searle's explanation that there is a form of coerced life in the institutionalized collective way of living is accepted in the article, and the further explanation is analyzed through his concept of bad faith, which is caused by the unreflective thinking by rational beings. (As he states, the 'can' is not necessarily followed by 'ought', though the latter causes the former.) But whether rule-following beings are making that decision rationally or not is the question. The criticism of Searle's explanation lies in the fact that:

a) The Searlean notion of social power is mostly Weberian (limited to agents, power over each other), but mainly cognitive and mental. In the Searlean account, coercion is possible when agents start internalizing the constitutive rules, minimizing the other-oriented actions that may benefit them (others). And,

b) The coerced collective behavior of an individual is unreflective and becomes an incoherent behavior too, as, on the one hand, rational beings are free to choose, yet on the other, they are bound by the imposed rules, which are either meaningless or unintelligible to them; in other words, the rules are coercive in nature.

There needs to be an explanation that either people are rational, i.e., choosing freely to act, or are irrational and not choosing freely. A third option is that in cases where people are confused between the choices, they hence have something beyond the deontological concerns to clear out the pertaining confusion. Searle doesn't significantly evaluate crucial nuances

and implications between the rational, irrational, and dilemma-based actions when he discusses institutionalized collective behavior. It is also the case that collective intentionality, behavioral representations, etc., may take a form or leave room for intrinsic coercive aspects through status function, power, meaning, etc. Moreover, in the Searlean scenario, a sense of compulsion through collective intentionality and power may exert an unintended coercive aspect, which the Searlean approach doesn't focus on much. In this case, Searle's approach appears incomplete and leaves an explanatory gap from a metaphysical and epistemological perspective.

References

Douglass, C. F. (2001). *Rationality in action* (Review of the book Rationality in action, by J. Searle). Cambridge: MIT Press.

Durkheim, É. (1898). Représentations individuelles et représentations collectives. *Revue de Métaphysique et de Morale, 6*(3), 273–302.

Schmid, H. B. (2003). Can brains in vats think as a team? *Philosophical Explorations, 6*(3), 201–217. https://doi.org/10.1080/10002003098538750

Searle, J. (1983). *Intentionality*. Cambridge University Press.

Searle, J. (1995). *The construction of social reality*. The Free Press.

Searle, J. (2001). *Rationality in action*. MIT Press.

Searle, J. (2006). Social ontology: Some basic principles. *Anthropological Theory, 6*(1), 12–29. https://doi.org/10.1177/1463499606061731

Searle, J. (2010). *Making the social world: The structure of human civilization*. Oxford University Press.

Smith, B., & Searle, J. (2003). The construction of social reality: An exchange. *American Journal of Economics and Sociology, 62*(2), 285–309.

Thompson, M. J. (2017). Collective intentionality, social domination, and reification. *Journal of Social Ontology, 3*(2), 207–229. https://doi.org/10.1515/jso-2016-0017

Wittgenstein, L. (2001). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell Publishing.

Balkan Journal of Philosophy (2025), vol. 17, issue 2, 199–213

RUNNING WILD: TEMPORALITY AT THE SPEED OF LIGHT

Renjie Li*

Abstract: This paper examines temporality in the information age, drawing on the philosophical works of Martin Heidegger, Bernard Stiegler, and Hubert Dreyfus. By re-investigating the analysis of the Being of equipment in *Being and Time*, the paper argues that Heidegger not only unwittingly foresees the future of technological planetarization but also touches on the question of 'real-time' in the use of tools. Retrospectively, Stiegler radically re-elaborates on this question through the figures of Prometheus and Epimetheus, who give the structure of temporality in the name of what he called tertiary retention. However, the paper will show that when one thinks of the determination posed by modern algorithms in machine learning, the question of the future becomes significant. What characterizes the power of intelligent machines is the predictability of the future. However, this is not the spontaneous and subjective anticipation of the individual, but rather leads to the degeneration of the synthesis of reason and various misunderstandings of intelligence. Therefore, the paper suggests rearticulating the light-speed temporal dimension, emphasizing its relevance to concrete technologies and its ongoing significance for cognition in the digital era.

Keywords: Temporality, Heidegger, Tertiary Retention, Algorithm, Artificial Intelligence.

1. Introduction

Digital technology poses a challenge to the conventional understanding of time. Today, we human beings are not only living with technical equipment, but rather we are living in the technological system. In contrast to the thriving industrialization of the 20^{th} century, the contemporary world is dominated by digital devices and sensors, moving towards *technological planetarization*. Special attention will be paid to the advent of the theory of cybernetics established by American mathematician Norbert Wiener in his book *Cybernetics: or Control and Communication in the Animal and the Machine* (1948). Wiener and his colleagues' establishment of cybernetics stands as one of the most significant developments of the 20^{th} century, aligning with Claude Shannon's information theory and thus becoming the foundation of modern computation. This is what scientists and philosophers traditionally understood as first-order cybernetics in the 1940s (David Maulén de los Reyes, 2024, p. 203).

We first want to clarify that the central point for cybernetics lies in adjustment and feedback between different units instead of imperative control.[1] Therefore, in this sense,

[1]The Chinese translation of cybernetics literally means "control theory", which is problematic to some extent. During the Second World War, Wiener was assigned by the government to work on the automatic artillery that was able to anticipate the flying trajectory of aerocrafts.

*Department of Philosophy and Religious Studies, University of Macau, Macau; yc37732@um.edu.mo

we may consider it a question of *guiding*. What is being guided is technological information. In another article "Traditional and Technological Language" (*Überlieferte Sprache und Technische Sprache*), Heidegger argued that modern machines characterized by cybernetic principles (i.e., technological-calculative principles) are transforming language as *saying* (to presence, *Vergegenwartigen*) to signals (Heidegger, 1998, p. 140).

By recognizing and computing the information transmitted in different units, signals can be conveyed and understandable not merely to living organisms but also interpretable by computers. Therefore, Wiener proposes that the development of cybernetics (first-order) enables automatic machines to assimilate the behavior of organisms and imitate their functions. Consequently, the former no longer undergoes traditional Newtonian time associated with mechanical movement, but instead shares the irreversible, biological Bergsonian time with organisms. Through this argument, Wiener claims that cybernetics overcomes the opposition between mechanism and vitalism (Bergson used the latter to object to mechanism) (Wiener, 2019, p. 63). The new operational logic in first-order cybernetics allows Heinz von Foerster as well as Niklas Luhmann to conceive an epistemology that can develop cybernetics into a system theory, and therefore, it unifies other disciplines and is compatible with the work of *social institutions*, meaning that "the observer enters the system by stipulating his own purpose" (Von Foerster, 2003, p. 286) in what he termed a *cybernetics of cybernetics*, that is, second-order cybernetics. For example, this project has been actualized in some Latin American countries on decentralizing computing management for the government, including Chile, Mexico, and Brazil (David Maulén de los Reyes, 2024, p. 229).

However, what is prominent today is not the theories that were coined by Wiener, Shannon, or Foerster. Rather, we see the proliferation of media platforms like X, TikTok, WeChat, and so forth. We might call these digital platforms in the name of third-order cybernetics. Bernard Stiegler acknowledges that digital platforms as such build the hyper-synchronization of temporality that integrates the time of consciousness (*Zeitbewusstsein*) across diverse time-space structures into a common consciousness. In other words, what characterizes digital platforms is the manipulation of individual consciousness. Retrospectively, we might find ourselves still confined within the condition of *Gestell* (enframing) that Martin Heidegger proposed in one of his lecture series, Das Gestell in 1949, where he questioned the essence of technology, and later published as "The Question Concerning Technology" (*Die Frage nach der Technik*, 1953). *Gestell* manifests itself as comprehensive calculability as well as unlimited accessibility to the standing reserves (*Bestand*), including various natural resources, that is to say, nature becomes supply energy. This is evident when considering the textual resources utilized by ChatGPT. Despite concrete advancements in information processing and programming, exemplified by companies like OpenAI and NVIDIA, questions persist regarding the temporality inherent in digital objects.

The paper argues that temporalization relies fundamentally on material support, particularly artificial constructs. This suggests that prior to 'my' present existence in the world, exteriorized technology already serves as a constitutive dimension for time.

Consequently, temporality varies in human perception under different conditions, such as between natural and digital objects. The importance of rearticulating temporality to light speed in digital technology lies in its role as a determinative force in our lives, rather than a

passive tool awaiting use. Therefore, this paper proposes to revisit Martin Heidegger's and Bernard Stiegler's discussions on technics per se and its relation to temporality.

On the one hand, Heidegger adopts an *onto-historical* approach to the essence of technology. Heidegger's approach ultimately returns to the question of the oblivion of Being (*Seinsvergessenheit*). The way Heidegger situated his ontological analysis of technology has often been criticized by respected philosophers, for example, Don Ihde, since frontier technoscience in current time is "post-Heidegger", which is beyond the categories of industrial technology in Heidegger's lifetime (Ihde, 2010, pp. 4-5). Therefore, Ihde implies that Heidegger's thoughts on technology will fail to read the concrete and empirical technological evolution. In the following section, the paper will show an alternative reading to argue that Heidegger's analysis of the Being of equipment in *Being and Time* (1927/2001) implies a prophecy of real-time modern technology.[2] On the other hand, Stiegler gives a more radical interpretation of 'time is *Dasein*' and its relation to historicity by unfolding the notions of *epiphylogenesis* and *tertiary retention*, arguing that it is only through the exteriorized artifacts that *Dasein* can find its past, that is, a new temporal dimension is opened. Thus, Stiegler partially refuses Heidegger's ontological difference, reconstructing it as the ontological difference between "the *who*" (Dasein) and "the *what*" (prosthetics) (Stiegler, 1998, p. 240). In the following pages, the paper endeavors to demonstrate the necessity of a profound analysis of the temporality of *Dasein* in connection with the examination of the Being of equipment in *Being and Time*. It is within this framework that we can attain a comprehensive understanding of temporality in the digital era and its significance for the continuity of thinking (*noesis*).

2. Between *technē* and *Gestell*: Equipment in *Being and Time*

Heidegger's notable "The Question Concerning Technology" (1949) provides an onto-historical perspective on technics, delineating the transformation from *technē* to *Gestell* and questioning the essence of technology. For Heidegger, the notion of *Gestell* represents his diagnosis of the mode of existence of modern technology as challenging (*Herausforderung*). In other words, modern technology possesses the capability of revealing but violently gathering all accessible resources and manifests itself as accurate calculability on *Bestand* based on modern science and experiments (Heidegger, 1977, p. 19).

Heidegger is focused on the more transcendent issue of manipulation that arises from a technological understanding of Being, rather than on concrete technologies. The danger arises from technological understanding becoming the only way of *thinking* – that is, the only way for the production of knowledge, which largely manifests as 'accelerated computing' in current times. In that case, a clearer way is presented when Heidegger foresees the power of cybernetics to coordinate all other disciplines at the time, including philosophy.

The end of philosophy proves to be the triumph of the manipulable arrangement of a scientific-technological world and social order proper to this world. The end of philosophy means: the beginning of the world-civilization based upon Western European thinking (Heidegger, 1993, p. 59).

[2]To describe the temporal condition of digital technology, I will use the terms "real-time" and "light speed" interchangeably.

Heidegger referred to the completion of Western philosophy as attributed to cybernetics because what originally characterizes philosophy is gaining true knowledge through reflection. And now, cybernetics shares the same function of reflection. Cybernetic principles have accelerated the technological planetarization of the world, while Heidegger observed: "... the epoch of the planetary lack of *mindfulness* (Besinnung)" (Heidegger, 2016, p. 38) – that is to say, we still do not know how to think. To counter this, Heidegger proposed the other beginning (*Der andre Anfang*), a project aimed at deconstructing technological totalization. This requires confronting danger itself, as new thinking is only possible when it remains *compatible* with technology but diverges from a purely technological understanding. After suspending the project of *Being and Time*, Heidegger began to explore the meaning of Sein through poetry and art (e.g., the question of ἀλήθεια). In his later work, Heidegger introduced the concept of the fourfold (*Geviert*) to examine the nature of objects.

Retrospectively, Heidegger claimed that the essence of technology is "by no means anything technological" (Heidegger, 1977, p. 12), emphasizing that his concern was not with the ontic dimension of technology but with its essence. Though Stiegler writes, "... or by all those who, perhaps imagining that they precede him (*Heidegger*), actually follow in his wake" (Stiegler, 1998, p. 225), Stiegler here means to challenge Heidegger's constraints on the metaphysical understanding of technics, which failed to truly grasp the themes of *Eigentlichkeit* and falling. Therefore, we may argue that Heidegger's concern needs to be reflected critically with respect to the concrete application of technology.

Heidegger himself re-emphasizes the *step-backwards* approach to *thinking* in the later "The Provenance of Art and the Destination of Thought" (1976). But back to where? Given that planetarization at the time was taking shape, any endeavors to revert to a primitive era would be ill-posed. For Heidegger, it is a journey back to the origin of European thinking (i.e., the ancient Greek) and to investigate what has been unthought, namely the *truth* that is concealed yet. In "The Origin of the Work of Art" (1935/36), Heidegger associated the meaning of "truth" with Greek *aletheia* (ἀλήθεια). *Aletheia* derives from the root *letha*, meaning forgotten. To confront the forgotten, *aletheia* demands presence (*Vergegenwartigen*).[3] *Aletheia* is to act out (Stiegler, 2011b, p. 223), an act of presenting something, which Heidegger sought to reinterpret in terms of *technē*. We may say that Heidegger worked out a meaning for *aletheia* in a way that even the ancient Greeks had not conceived.

For Heidegger, *technē* and *Gestell* both represent a mode of revealing, but they unfold themselves at different scales. *Technē* refers to a revealing of what does not yet present to us to the presence, that is, a bringing-forth (*hervorbringen*), allowing things to be seen, by which *aletheia* can be unveiled (Heidegger, 1977, p. 10). The unconcealment for truth or *aletheia* depends on the formation, that is, clearing (*Lichtung*) or appearing. Heidegger clarifies this process of "coming-into-being" by referring to Aristotle's four causes when it comes to the case of making a chalice (Heidegger, 1977, pp. 11-12). This may lead us to the analysis of the Being of equipment/tools that Heidegger already touched on in *Being and Time*, beginning from §14 to §18. In these sections, Heidegger locates the analysis of equipment in the

[3]Because truth is constantly forgotten, *anamnesis* becomes, in Stiegler's terms, the process of recollecting forgotten truth. I will explain this in the third part of this paper.

phenomenological world to provide the horizon to understand human existence as "being-in-the-world", and re-articulate *concerned encountering* in terms of equipment-referential structures over the knowing through theoretical knowledge.

Equipment presupposes *Dasein*'s (human beings as "being-in-the-world" and capable of apprehending their Being) understanding of tools, for which in this way one can *use* tools that conform to their *telos* (in-order-to), that is, the most proper way of *appropriating* equipment (Heidegger, 2001, pp. 97-98): for example, my a priori interpretation of hammers towards *hammer-ing*. In other words, equipment or tools stand no longer in front of me as mere objects, but rather each piece of equipment bears *the range of using-to-do* that precedes the theoretical understanding to fit its usability, serviceability, and so forth. In this sense, each tool operates within an equipmental totality, wherein no tool exists in isolation, but only in the context of manifold assignments, meaning that all tools are internally related to rendering the context of totality (Zimmerman, 1990, p. 139), in which the Being of equipment can be presented, also the meaningful world can be revealed. Heidegger writes, "the relational totality of this signifying we call 'significance' (*Bedeutsamkeit*). This is what makes up the structure of the world" (Heidegger 2001, p. 120). Therefore, only a "relatively autonomous local context" (Dreyfus, 1984, p. 31) enables not only the being of equipment but all entities to manifest themselves. Here, we discern the mode of existence of equipment as "readiness-to-hand" (*Zuhandenheit*) (Heidegger, 2001, p. 101). The readiness-to-hand of equipment also means that tools *withdraw* from utilization and thus become invisible, as water is to fish. Nevertheless, we shall argue that it is precisely the absence of equipment that reveals the *already-there* world that pre-*involves* the concern of *Dasein* with tools that signify "in-order-to".

Heidegger contends that it is only through the occurrence of damage or breakdown of tools in practical activities that we may find the complex referential context in which they are involved. The paper contends that this *negative approach* precisely refers to the *operational limitation* of these simple tools in *Being and Time, which at least unwittingly implies a discourse on technical evolution until the condition of* Gestell. The breakdown of the equipment means that it will lose its serviceability, or we might say *performativity*, in its depletion. The dominant philosophical interpretations of *Zuhandenheit* do not address the question of depletion of equipment, nor is Heidegger's original intention in *Being and Time*. However, I argue that to maintain performativity, equipment must rely on external structures to achieve a new stage of concretization. This process increases structural and functional complexity, countering the inevitable rise in entropy – a challenge faced by all physical beings, as explained by physics. The notion of entropy was also adopted by Stiegler, arguing that consumerism has become the dominant standard in digital society, thus accelerating entropy in a thermodynamic sense (Irwin, 2020, p. 372). Therefore, it is important to concretely articulate this question to reveal the connection between the depletion of technical objects/tools and the technical evolution toward *Gestell*.

One effective example is the Graphics Processing Unit (GPU). With its large number of transistors, GPUs effectively improve computation in deep learning, thus supporting the proliferation of artificial intelligence. However, the increasing number of transistors within the GPU also means higher power consumption and heat generation, which ultimately degrades performance – a manifestation of entropy. To cope with the challenge of heat generation,

the GPU must adopt different cooling technologies (e.g., air and water cooling) to reduce power consumption while at the same time improving performance. This coupling marks a new stage of concretization, measurable through the metric of performance per watt (PPW). Tools, in this sense, are not merely beings without dynamics.[4] Then, the question of relational totality becomes clear. No matter if a tool is in the reticulation of equipment or the encounter with *Dasein* in an ontic sense, there is always a demand to couple with other entities, including with *Dasein*. However, with the concretization of technical beings, first, they become sophisticated, *narrow*, and special to certain groups who possess the knowledge of concrete technologies, which consequently results in the alienation between technical individuals and humans. Second, during the process of becoming sophisticated, they become vulnerable at the same time. Each concretization towards perfection requires coherence among different units; as Gilbert Simondon proposes, the aporia for technology is how to allow different functions to converge within a shared structure (Simondon, 2017, p. 28). This convergence entails the narrowing of the scope of use, directing the evolution of equipment, and transforming the relational totality into an increasingly perfected closed system at an ontic level. Ultimately, this process culminates in the world evolving into a vast technological system capable of mobilizing global resources, known as *Ge-stell*.

Though Heidegger asserts that *Gestell* represents the essence of modern technology, which is "nothing technological", this understanding of technology is nonetheless rooted in the concretization and operation of tools, which attempt to confront entropy through *real-time adjustments*. Returning to the example of the GPU, its temperature, power consumption, and cooling mechanisms change every second. Today, the condition of *Gestell* largely manifests itself as the planetary digitalization supported by computing programs in various forms, such as the Internet, digital maps, and cable networks (Kisiel, 2014, p. 138). More critically, we now face the condition of the "IT *Gestell*" (Irwin, 2020, p. 371). Heidegger did not critically discuss real-time adjustment, because, for him, *Zuhandenheit* is the strategy to deconstruct the dominance of propositional knowledge in Western metaphysics, while the depletion of tools is only an exceptional situation to make tools present-to-hand. For Stiegler, the question of real-time is something *unthought* in Heidegger's thought; at least, it has not been radically emphasized as we have stated.

3. Now Come Fire, or Artificial Organs

Heidegger's analysis of technics is largely through *ontologicalized* equipment *per se*, which is also compatible with what he discerns as the task of thinking after cybernetics. Still, it is a transcendental task. On the contrary, Stiegler conducts an *a-transcendental* task, which means that what is transcendental is vulnerable to being conditioned (Hui, 2022, p. 90). Conditioned by what? To address this project, we argue that the question of the artificial is not something that can be overlooked.

For Stiegler, the Being-question that Heidegger discerns is situated on another question of *forgetting*, namely, the forgetting of technics. The forgetting of technics arises from Stiegler's reading of the myth of Prometheus in Plato's *Protagoras*. When Epimetheus made the first fault that he *forgot* to distribute qualities (i.e., skills) to mortals (humans), then

[4]For Stiegler, that is the reason technical beings have been devalued since the beginning of the history of philosophy. See *Technics and Time 1*. General introduction: 1.

Prometheus stole *fire* from Olympus to supplement humans, doubling the fault (Dillet and Stiegler, 2017, p. 48). Fire, as technics, constitutes the supplement to humans (i.e., *prosthesis*) as the origin of humans is a *default*, but it is a necessary default for opening the history of being of *Dasein*. This necessary default echoes what Heidegger analyzes as "Being-guilt" since *Dasein* always "has debts to someone" and is responsible for its own being, that is, being-towards-death (Heidegger, 2001, pp. 327–225; Wang, 2022, p. 144). In other words, the temporal mode of being of *Dasein* anticipates its death, which manifests as the temporality of ecstasies. Heidegger gives priority to 'my' death, which is the most authentic, is one's ownmost (Heidegger, 2001, p. 250). However, the temporality of ecstasies of anticipating one's death requires itself to be *historical* because it is only through the accumulation of history that one can conceive death in an indeterminate way. Stiegler argues that this anticipation demands a repetitive return to *Dasein*'s past and its present, and meanwhile, it is "a past that is not its past" (Stiegler, 1998, p. 232). It is a past I never lived, but without this past, I would not exist as well. Therefore, in this context, *Dasein* can either interpret its being through an ordinary understanding of itself in the tradition or the inherent past that is *not* its past "on the basis of its possibility" (Stiegler, 1998, p. 5). Death is always anticipated in the world through the death of others since the already-there world is constituted by *Dasein* being with others, which means a history of being always prior to *Dasein*. *Dasein* is *its* past and anticipates from it, and consequently, is always in being of *has-to-be*. Therefore, Stiegler contends that this can only be possible through prosthesis in terms of Epimetheus (Stiegler, 1998, pp. 232-233), who constitutes the necessary default for technics, making the condition of historic-technical repetition to *Dasein* possible.

Technics constitutes 'what' is capable of retaining in the *différance* (with an "a") of consciousness and allows the memory to have the right to be summoned regardless of the passing of time; that is *tertiary retention*. Here, we may first go back to how Stiegler's concept of tertiary retention responds to Husserl's position on primary and secondary retention. Stiegler himself acknowledges the difference between primary and secondary retention to recognize mental activities, but he also contends that it is precisely through the function of tertiary retention that the two can work properly (Ross, 2023, p. 17). Primary retention refers to the 'originary impression' but is always in the process of no longer, as "I am now hearing the melody" (Stiegler, 2011a, p. 14), the 'now' that is not fully passed away. Secondary retention is a recollection of what has completely passed away and that now is not there, namely, memory. Stiegler thinks of tertiary retention in relation to Husserl's "image-consciousness" (Stiegler, 1998, p. 248; 2011a, pp. 14-15). Image-consciousness is possible through image objects like paintings or pictures. However, Stiegler contends that whether the image object is actually being apprehended or only imagined (for example, regardless of whether it is a real tree perceived by the subject or is the experience imagined by the subject), both experiences carry the task for activating "image-consciousness" (Stiegler, 2011a, p. 20). Stiegler then proposes that perception is never the pure perception of consciousness. For Stiegler, perception comes from primary retention, which is selected by secondary retention (in this sense, the sedimentation of primary retention now becomes secondary but is not identical to the genuine experience that is already lost, the reason for which secondary retention is always *re*-presentation). In Stiegler's words, perception is not merely confined to primary retention because consciousness is not able to repeat the perception identically, as pure repetition means no difference, which indicates there is no longer the flow of time.

Then, perception in the repetitive process already contains a certain extent of imagination due to the *default* in perception (Stiegler, 2011a, p. 19). For example, in our repetitive listening of the same melody, we will find that the experience of listening varies because the selection to primary retention is also transformed. This transformation comes from primal listening as the basis. This is also Stiegler's criticism of Husserl, who differentiates but also *opposes* perception with recollection/imagination (Stiegler, 2011a, p. 16, 21).

More importantly, Stiegler emphasizes that the selection (to primary retention) of secondary retention is conditioned by tertiary retention. Tertiary retention possesses the competence of what Stiegler called "orthothetic" (*orthothétique*) (Stiegler, 1998, p. 4, pp. 274-275), that is, an accurate recording of temporal objects such as music or film. Stiegler mentions *writing* as a technique that allows the transmission of truth from generation to generation in *Technics and Time Vol. 1*. For example, our interpretation of Heidegger depends on the historical texts that faithfully recorded his thoughts. The position of this supplement relates to Stiegler's reading of the theory of individuation of French philosopher Gilbert Simondon, which is also the key to understanding Stiegler's thought. Simondon uses the process of individuation to depict the *becoming* and mutation of technical objects. We might understand this process through an example of crystallization (this is Simondon's frequently used example). Consider an over-saturated salt solution where the internal tension or *disparation* within the solution reaches its limit. The introduction of minimal information, for example, heating, can trigger crystallization, that is, the *supplement* in our case. Then, crystals will emerge and eventually reach a new metastable state. This represents the process of individuation on a physical level. Stiegler borrows this term from Simondon and tries to reconceptualize it as *elevating*.[5] Tertiary retention as the supplement (e.g., writing) is what is capable of elevating humans, or *Dasein*, Heidegger would say, to a plane of consistency for new thinking or understanding to emerge, with its temporality that supports the temporal flux of reasoning. Like what Stiegler himself often explains, in *Meno*, the young man recollects the truth by drawing geometric figures on the sand, which retains his reasoning (Dillet and Stiegler, 2017, pp. 51–52).

However, the problem is that this supplement can, on the one hand, facilitate the new structure to emerge in the Simondon sense, but, on the other hand, can simultaneously deconstruct our reasoning process, that is, *dis-individuation*. This deconstruction manifests as the *degeneration* of pure *understanding* in a Kantian sense; the capacity to synthesize now is automated and has been replaced by big data, or in other words, artificial reason. It is in the *Technics and Time Vol. 3* that Stiegler argues that the Industrial Revolution gave rise to new forms of tertiary retention, capable of reshaping knowledge production and influencing unconscious behaviors directly, bypassing conscious awareness, implying that the manipulation of desire or *phantasia* would be more efficient in guiding consumption (Stiegler, 2011b, p. 225). To this point, desire is being reduced to a drive in the cultural industry. Desire, in Stiegler's work, means a necessary investment of time, that is, *attention*. The short circuit to desire leads to a drive that refers to instinct; for example, I eat when I feel

[5] "... is to individuate, without which one risks disindividuating, like the herds described by Nietzsche, or what Heidegger calls *Verfallen*, the fall in which one loses his or her self-ness or authenticity (*Eigenlichkeit*)." See: Hui, Yuk. (2024). Bernard Stiegler, the mystagogue. *Philosophy Today* 68 (3): 585–586. Regarding the conceptualization of elevating and Stiegler's interpretation of the myth of Persephone, see Stiegler, Bernard. 1996. Persephone, Oedipus, Epimetheus. *Technema*, translated by Richard Beardsworth, 3: 97–112.

hungry. Therefore, for Stiegler, the artificial participation of tertiary retention can either "open up new possibilities *or* undermine the capacity for remembering and knowing" (Ross, 2023, p. 17). The situation of undermining the capacity for knowing is what Stiegler called *proletarianization*. This concept implies the loss of both theoretical knowledge and the knowledge of how to live (*savoir-vivre*), and besides, one cannot apply his/her knowledge in hyper-control societies (Stiegler, 2022, p. 279).

This is why Stiegler suggests we rethink Aristotle's theory of the four causes, especially *causa efficiens* (Stiegler, 2022, p. 280). The theory of the four causes relates to the integration of form and matter, for what apparently is being *digitalized* in our time. With the emergence of new kinds of tertiary retention, knowledge is *encoded* and *programmed* into information. Knowledge in the form of information means that it is conceived to be exchanged, and thus gives priority to efficiency. For Stiegler, the hegemony of the efficient cause that comes from its calculability ultimately affects the relation to final causes by putting them into operation (Stiegler, 2022, pp. 275-276). Given that the progress of coding and production is largely actualized in the research of generative artificial intelligence (Generative AI), exemplified in ChatGPT, I shall argue that the *references to time and space* have been changed. Therefore, I propose to go back to Heidegger's analysis of significance with the reading of Hubert Dreyfus to unpack the continued relevance to temporality at the speed of light.

4. The Aporia of Temporality in Algorithms

Generative AI like ChatGPT and Midjourney has raised extensive discussion in recent years. Beneath the fantasies of God-like superintelligence, I argue that AI manifests as a concrete technology – specifically, as an artificial neural network. The efficiency of this technology is highly related to, first, the progress of hardware like GPUs, and second, machine learning techniques, namely algorithms. The efficiency of this technology is closely tied to, first, advancements in hardware like GPUs, and second, machine learning techniques, particularly algorithms. Retrospectively, debates on the construction of AI programs can be traced back to the 1970s. In *What Computers Can't Do* (1972), Dreyfus suggests that calculative thinking emerges early in the history of philosophy, as seen when Socrates queries Euthyphro about justifying piety, requiring knowledge of all scenarios to establish a standard (we might say this is the earliest version of the method of exhaustion) (Dreyfus, 1972, p. 15). This thinking approach as an efficient program eventually becomes the central aim of metaphysics in apprehending reality, a journey spanning from Plato to Hobbes, Leibniz, and Norbert Wiener, and ultimately leading to the flourishing of artificial intelligence. Dreyfus criticizes the research on AI, especially the early work on problem-solving of Alan Newell, Hubert Simon, and Marvin Minsky, for dedicating substantial effort to recognizing significance (*Bedeutsamkeit*) in the real world with facts and representational symbols, aiming to prove that computers can perform tasks beyond mere calculation. However, Dreyfus highlights the risk of an infinite regression within these cognitive frameworks, as continually adding parameters is necessary to encompass all possible situations. Consequently, he argues that AI research focused on cognitive simulation was a degenerative project, falling far short of human intelligence. In other words, conventional approaches to AI research effectively address specific issues with limited parameters (frames), yet they fail to generate common-sense knowledge due to erroneous philosophical assumptions (Dreyfus, 1972). Dreyfus then subtly

points out that the problem lies in the way of knowing. The transformation of the underlying philosophical assumption suggests that only when we properly formulate the question of AI can we then adequately reveal the question of temporality in reticulated artificial intelligence (including Generative AI).

> Minsky, unaware of Heidegger's critique, was convinced that representing a few million facts about objects, including their functions, would solve what had come to be called the commonsense knowledge problem. It seemed to me, however, that the deep problem wasn't storing millions of facts; it was knowing which facts were relevant in any given situation. [...] If the computer is running a representation of the current state of the world and something in the world changes, how does the program determine which of its represented facts can be assumed to have stayed the same, and which would have to be updated? (Dreyfus, 2007, p.1138).

What is Heidegger's critique, and why is it relevant here? We might go back to Heidegger's discourses on signs and significance in *Being and Time*. As we have shown in the second section, Heidegger claims the priority of practicing (i.e., ready-to-hand) rather than contemplating cognition. That is to say, to understand a chair or stone, for example, does not mean to collect all its values or facts, but rather, one needs to realize its relation to other equipment as well as the whole of related human activities, which I shall not repeat here. The critical point to note is that attempting to analyze all human activities and encode them into formal logic is illusory. Firstly, phenomena such as mood states, like embarrassment or suicidal ideation, often lack clear manifestations. Secondly, our resolution, as Heidegger terms it, is not entirely predictable; if it were, it would cease to be a resolution. Furthermore, the resolution is not a one-time event, which means that we constantly conduct the determination all the time through the concerned engagement with tools to apprehend the worldhood of the world, which involves the question of the temporal dimension of real-time, that is, without delay. What is revolutionary is that the real-time response in the use of tools is precisely the nature of contemporary reticulated artificial intelligence (e.g., ChatGPT normally responds to prompts within seconds), including all digital platforms, Google, ChatGPT, and so forth, and thereby reforms the concept of time and space.

Retrospectively, Dreyfus's critique emphasizes the already-there world that precedes any cognition, advocating for AI research to transcend the reliance on formal representation. Later developments in machine learning in connectionism and Walter Freeman's neurodynamics align with this trajectory (Hui, 2021a, pp. 238–240). However, we might say that Dreyfus focuses on the hermeneutics of the tools and the world, but he does *not* pay much attention to the question of algorithms. For Hui, what is concealed under this issue is the question of recursivity (Hui, 2021a, p. 241).[6] Our aim here is not to conceive how to build an AI program; rather, our main concern is to introduce the aporia of temporality posed by algorithms, a problem that was not prominent in Dreyfus's time. If we can say that the primordial temporality of *Dasein* makes historicity possible, and Heidegger is correct that *Dasein* cannot be understood merely through the temporal dimension of the present,

[6]Yuk Hui develops the notion of recursivity (in his book *Recursivity and Contingency*) to refer to the movement that the subject constantly comes back to itself and to recognize itself. For a more detailed discussion on recursive algorithms, please refer to *On the Existence of Digital Objects* (2016).

Stiegler then points out what Heidegger overlooked is tertiary retention that supports the repetition of the past for *Dasein* to access, so that of which can stem from there. Tertiary retention is generally understood as the exteriorization of memory, and the new forms it takes in the information age are digital. Algorithms not only seek optimal solutions but also absorb mutations (i.e., contingent variations) during the iterative process. For example, until 2013, Google had already implemented 200 parameters in the search algorithm, shaping a digital landscape *tailored* to each user.[7] Therefore, I argue that algorithms create a cyberspace uniquely constructed for each user (I will return to this point later when discussing the question of the future). Retrospectively, digital platforms, in the context of third-order cybernetics (as we discussed earlier), stack the flux of users' consciousness, analyzing, relocating, and then listing all possible hyperlinks on the user interface. This process allows one's existence to transcend the conventional temporal dimension of 'here and now', as 'my' conscious activity can simultaneously manifest across various data networks on the internet. What changes here is the pace of time. As we have outlined the transformation of the condition of *Gestell* from Heidegger's time to the information age, we can see that communication no longer relies on postal services, but rather on fiber optic networks supported by integrated telecommunications infrastructure. As Stiegler pointed out, data transmission within global fiber optic networks occurs four million times faster than the transmission of biological signals in the human body (Stiegler, 2018, p. 84; Irwin, 2020, p. 372), reinforcing global synchronization at the speed of light. This was exemplified when Derrida arrived in Shanghai on the 11^{th} of September 2001 and learned of the Twin Towers disaster on the same day at midnight through televised images (Borradori, 2004, p. 110). Aside from the event itself, Derrida received the news at roughly the same time from the other side of the world. In other words, real-time communication without delay allows the "who" to be present anywhere at the same time.

Beyond the synchronization created by real-time communications in digital platforms, real-time interactions also involve the computation of the flux of consciousness from various users, at the speed of light. Consciousness only becomes computable through two prerequisites: first, it must be artificial or material, and second, there is the necessity of objectification. We have discussed how tertiary retention actively participates in the structure of understanding through Stiegler's reading of Heidegger; I now emphasize the second point. The implications of real-time technology on consciousness led Stiegler to re-examine Immanuel Kant's *Critique of Pure Reason*, focusing on the problem of transcendental imagination. For Stiegler, Kant could not recognize:

> Consciousness can only become *self*-consciousness when it can be externalized, objectivized as traces through which at the same time it becomes accessible to other consciousness (Stiegler, 2011a, p. 47).

Stiegler's text highlights the importance of tertiary retention in the process of uniting the flux of consciousness. It also suggests that what one is doing in the present was experienced by *others* at a different pace of time, through interobjective reorganization. By "interobjective," I refer to the concept developed by Hui in *On the Existence of Digital*

[7] See Dean, B. *Google's 200 Ranking Factors: The Complete List.* Retrieved March 24, 2024, from `https://backlinko.com/google-ranking-factors`.

Objects. According to Hui, interobjectivity has two meanings: first, it refers to the materialization of all sorts of relations of objects, and second, materialized interobjectivity creates its own milieu but can also be transformed into contacts between objects (Hui, 2016, p. 160). Retrospectively, I have mentioned how algorithms measure individual preferences and *tailor* unique cyberspaces to the needs and preferences of users. This may seem subjective, akin to how time is inherently tied to *Dasein*'s subjectivity in Heidegger's analysis. However, cyberspace created by algorithms is neither independent nor subjective. Through the universality of markup languages, data can be shared and interacted with. In this sense, for example, the recommendation shown in my user interface not only reflects traces of my past but also integrates others' conscious minds who share similar preferences. More importantly, others may have already made decisions regarding a possible recommendation and thus influenced the final recommendation presented to me. Therefore, the future (i.e., the next move) of the subject is projected by its own past and the past of others who have already experienced it, which occurs due to the continuous re-organization and emergence of data clusters across various data servers and digital traces left by different individuals over time. In other words, one's subjective projection of future events can be both *post-proper* (like in normal decision-making) to my awareness and simultaneously *prior* to it (determined by algorithms). As Hui reiterated this point:

> My emphasis, as mentioned, will be on the future. I would like to put my arguments in a simple proposition: our past and present are becoming more and more determined by future events that have not yet happened but are paradoxically already there. [...] a form of anticipation that is no longer my own subjective projection but rather a projection that is imposed on me and proceeds ahead of me (Hui, 2021b, p. 79).

Algorithms preemptively anticipate the individual's future that is *not subjective* from the analysis of the past and the present that are retained by tertiary retention, and then are able to determine the present and edit the consciousness of the past in reverse. For Yann LeCun, "prediction is the essence of intelligence."[8] The recursivity of deep-learning algorithms precisely shows predictability; it predicts expected outcomes through learning from predicted errors. In the example of searching, the design of the search engine is against the indeterminacy beneath the action of searching. For example, I search for the answer on Google because I do not know the answer, while what Google is trying to do is to build the answer *prior to* the occurrence of the question, a project which is known as "Google Now" or "Google Assistant". Therefore, the question of the future is not the problem of algorithms, or we might say, it is the capability to predict the future that makes them so powerful. It is analogous to the debate that arises in the domain of artistic creation: some may still argue that the work generated by AI cannot be considered real artwork since the machine does not understand the question of the infinite, such as the quality of depth for Paul Cézanne, and colors for Claude Monet. However, the power of AI lies in its ability to inscribe the infinite with finite forms (Hui, 2023, p. 6), just as algorithms predict the future through recursive computation and ultimately marginalize subjective anticipation (i.e., without specific interaction with the subject). In other words, this situation implies

[8] See X. Yann LeCun (@ylecunn). Retrieved 2017, from https://twitter.com/ylecun/status/864463163135324165.

the precision of the prediction. On the contrary, *Dasein*'s anticipation of the future, for example, death, is indeterminacy as the knowledge of death always withdraws itself.

For Heidegger, *Dasein*'s anticipation of the future is essentially an anticipation of death (*Vorlaufen zum Tode*). It is through this process that *Dasein* can realize its future possibilities by fixing the meaning of time. Here, we see Heidegger's analysis of inauthenticity (or, *Unentschlossenes Verfallen*) remains relevant. The nature of media platforms requires us to stay online, interact, and produce data. When we lose ourselves in social media, then we fail to recognize our finitude (Coeckelbergh, 2022, p. 8). Algorithms transform possible future events into the inauthentic presenting (*Gegenwärtigen*), where our consciousness is trapped in a state of 'disorientation'. *Dasein*'s anticipation (*Vorlaufen*) is largely re-modified by algorithms and becomes mere awaiting (*Gewärtigen*), an inauthentic mode of being as *Verfallen* that Heidegger described, and risks becoming the 'herds' for Nietzsche (Hui, 2024, p. 585). In this way, the predictability of algorithms rather intensifies the phenomenon of *das Man*.

Suppose the question of the future/infinite is not the case in reticulated artificial intelligence. How can we think of our future as one which escapes the over-determination that comes from the temporal dimension at the speed of light (the future knitted by algorithms)? It seems like the over-determination of this temporality is already embedded in the myth of Prometheus and Epimetheus, the figures that give the dimension of temporality. Fire is never intrinsic to *Dasein*, nor is it intrinsic to the gods; unlike the thunderbolt to Zeus, fire is fickle and can easily perish (Stiegler, 1998, p. 286). Fire is a stranger to *Dasein*, and thus bears the danger of being wild. The contrast between the power of fire to powerless mortals is likened to that of the preemptive algorithm to reason. However, the mortality of *Dasein* gives the meaning that it is perpetually incomplete. It is the being that always *has-to-be*. *Dasein* is never a state to be determined, but rather it implies a process of becoming, and encountering what is unknown. The disquieting consciousness when reason encounters the unknown (e.g., death, the explosion of the earth) is not merely a byproduct of rationality, but it characterizes the nature of reason per se. The meaning horizon of the existence of *Dasein* is to constantly cultivate itself toward a new phase of thinking, or what Stiegler called to *individuate*:

> The earth then becomes the place where we contemplate the sky *but by writing out what happens*. And the underground where the spirits of the dead were held disappeared: the hell of Hades, which also gives its name to the Greek aidôs, is no longer beneath the earth. There is no longer anything beneath the earth. Everything is played out between the sky and the earth. (Stiegler, 2006, p.336)[9]

Intelligence is currently defined by the precision of calculation, and in this sense, superintelligence is nearer.[10] However, this mode of thinking, according to Stiegler, withdraws

[9]Stiegler, Bernard. (2006). Chute et élévation. La politique de Simondon. *Revue philosophique de la France et de l'étranger* 131 (3): 336. Cited from Hui, Yuk. (2024). Bernard Stiegler, the mystagogue. *Philosophy Today* 68 (3): 596.

[10]I personally believe this debatable understanding also leads to a series of ethical debates in the research of AI in China, while simultaneously overlooking the importance of the artificial, despite the country's growing competitiveness in global AI research. We might want to acknowledge the contributions of contemporary

myths and mystagogy, thereby eliminating the potential for *surprise*, or "super-prendere" in Latin, which means that it transcends what is foresighted. Retrospectively, human intelligence extends beyond mere *recursive* thought (a step that already surpasses Dreyfus's critique); it encompasses the capacity for creation. It is through creation that the subject reveals its presence and the past, which *spontaneously* perceives the possible future and gives meaning to the horizon of one's existence. Art emerges as a particularly fertile ground for this endeavor, not merely as a discipline, but as a novel form of fragmented knowledge. While reticulated artificial intelligence reshapes rationality into discrete forms, it also opens avenues for contemplating fragmentation within the system and striving to transcend it; that is, many questions are not able to be answered in a universal form (e.g., what is the work of art?), meaning that the resolution must be multi-dimensional. This results in epistemological fragmentation through confronting technology. The fragmentation of reason results in diverse parallel units capable of comprehending knowledge independently of algorithmic/rational automation rather than adhering to the system's mendacious promise of perceiving them as mere *parts*. This new mode of thinking enables *Dasein* to philosophize about the unknown while avoiding assimilation by computation. Similar to how a painting is never truly complete for an artist, *Dasein* remains in a constant state of becoming.

Acknowledgments

I would like to thank the editors and two anonymous reviewers for their insightful and constructive feedback, which significantly improved the paper's presentation.

References

Borradori, G. (2004). *Philosophy in a time of terror: Dialogues with Jürgen Habermas and Jacques Derrida*. Chicago: University of Chicago Press.

Coeckelbergh, M. (2022). *Digital technologies, temporality and the politics of co-existence*. London: Palgrave Macmillan.

Dreyfus, H. (1972). *What computers can't do: A critique of artificial reason*. [1st ed.]. New York: Harper & Row.

Dreyfus, H. (1984). Between technē and technology: The ambiguous place of equipment in *Being and Time*. *Tulane Studies in Philosophy, 32*, 23–35.

Dreyfus, H. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence*, 171(18), 1137–60. https://doi.org/10.1016/j.artint.2007.10.012.

Dillet, B. & Stiegler, B. (2017). *Philosophising by accident: Interviews with Elie During*. 1st ed. Edinburgh: Edinburgh University Press.

Heidegger, M. (2001). *Being and Time*. (John Macquarrie and Edward Robinson, Trans.). Oxford: Blackwell.

Heidegger, M. (1977). *The question concerning technology and other essays*. (William Lovitt, Trans.). New York: HarperCollins.

Heidegger, M. (1993). The end of philosophy and the task of thinking. In *Basic Writings*. (David Farell Krell, Trans.). New York: HarperCollins.

Heidegger, M. (2004). The origin of the work of art. In *Off the Beaten Track* (Julian Young and Kenneth Haynes, Trans.). Cambridge: Cambridge University Press.

Heidegger, M. (2013). The provenance of art and the destination of thought (1967). *Journal of the British Society for Phenomenology*, 44(2), 119-28. https://doi.org/10.1080/00071773.2013.11006794.

thinkers to the study of technics, including Bernard Stiegler, Don Ihde, Bruno Latour, Yuk Hui, Pieter Lemmens, and Anna Longo, among others.

Heidegger, M. (2016). *Mindfulness*. (Parvis Emad and Thomas Kalary, Trans.). London: Bloomsbury Academic.

Hui, Y. (2019). *Recursivity and contingency*. London: Rowman and Littlefield.

Hui, Y. (2020). For a cosmotechnical event: In honor of Don Ihde and Bernard Stiegler. In G. Miller & A. Shew (Eds.). *Reimagining Philosophy and Technology, Reinventing Ihde* (pp. 87–102). Switzerland: Springer International Publishing.

Hui, Y. (2021a). *Art and cosmotechnics*. Minnesota: University of Minnesota Press.

Hui, Y. (2021b). Problems of temporality in the digital epoch. In Axel Volmar and Kyle Stine (Eds.). *Media Infrastructures and the Politics of Digital Time: Essays on Hardwired Temporalities* (pp. 77–87). Amsterdam: Amsterdam University Press.

Hui, Y. (2023). Imagination and the infinite – A critique of artificial imagination. *Balkan Journal of Philosophy* 15(1), 5–12.

Hui, Y. (2024). Bernard Stiegler, the mystagogue. *Philosophy Today* 68(3), 585–600.

Ihde, D. (2010). *Heidegger's technologies. Postphenomenological perspectives*. New York: Fordham University Press.

Irwin, R. (2020). Heidegger and Stiegler on failure and technology. *Educational Philosophy and Theory*, 52(4), 361–375.

Kisiel, T. (2014). Heidegger and our twenty-first century experience of Ge-Stell. In B. Babich and D. Ginev (Eds.). *The Multidimensionality of Hermeneutic Phenomenology* (pp. 137–151). Springer.

Maulén de los Reyes, D. (2024). Why did cybernetics disappear from Latin America? An incomplete timeline. In Yuk Hui (Eds.). *Cybernetics for the 21st Century: Vol. 1 Epistemological Reconstruction* (pp. 203–235). Hong Kong: Hanart Press.

Ross, D. (2023). Too soon too late: A pretext for a recurrence of Bernard Stiegler. *Media Theory*, 7(2), 01–36.

Simondon, G. (2017). *On the mode of existence of technical objects*. Minneapolis: Univocal Publishing.

Stiegler, B. (1996). Persephone, Oedipus, Epimetheus. (Richard Beardsworth, Trans.) *Technema, 3*. 97–112.

Stiegler, B. (1998). *Technics and time 1: The fault of Epimetheus*. Stanford: Stanford University Press.

Stiegler, B. (2011a). *Technics and time 3: Cinematic time and the question of malaise*. Stanford: Stanford University Press.

Stiegler, B. (2011b). The tongue of the eye: What 'art history' means. In Jacques Khalip and Robert Mitchell (Eds.), *Releasing the Image: From Literature to New Media* (pp. 222–235). Stanford: Stanford University Press.

Stiegler, B. (2022). The ordeal of truth: Causes and quasi-causes in the entropocene. *Foundations of Science, 27*(1), 271–280. https://doi.org/10.1007/s10699-020-09736-3.

Von Foerster, H. (2003). Cybernetics of cybernetics. In *Understanding Understanding*. Springer, New York, NY. https://doi.org/10.1007/0-387-21722-3_13.

Wiener, N. (2019). *Cybernetics: Or, control and communication in the animal and the machine*. MIT Press.

Wang, Q. (2022). Technics and time: From Heidegger to Stiegler [技術與時間：從海德格爾到斯蒂格勒]. *Philosophical Analysis* 13(5), 142–150.

Zimmerman, M. (1990). *Heidegger's confrontation with modernity: Technology, politics, and art*. Bloomington: Indiana University Press.

Balkan Journal of Philosophy (2025), vol. 17, issue 2, 214–227

A CRITICAL RECONSIDERATION OF THE ARGUMENTS OF SOCRATES FOR POLITICAL OBLIGATION

Ndukaku Okorie*

Abidemi Israel Ogunyomi**

Abstract: In one of Plato's major dialogues, *The Crito*, Socrates, as the leading protagonist, addressed one of the major questions and concerns of political philosophy. The question is: what should be the relationship between the state and citizens? This question raises the problem of political obligation, which is about the moral requirement or duty of citizens to obey the laws and directives of the state, for the first time in the history of philosophy. In defense of political obligation (the duty of citizens to obey the laws of their state) in the dialogue, Socrates demonstrated his respect for Athenian law in his refusal to flee his death sentence. His arguments for political obligation can be summed up in three categories, which are interconnected: first is the argument from the supremacy of the state; second is the argument from gratitude; and third, is the argument from contract. These arguments, for Socrates, justify the total obedience of an individual to the laws of his state. However, different philosophers throughout the epochs that succeeded Socrates have reacted to some of the issues raised by Socrates. Nevertheless, the question of political obligation remains inconclusively discussed in political philosophy; hence, it is an issue that is relevant in contemporary times. Against this background, we retrace the arguments as originally formulated by Socrates, re-examine them in light of the responses of other political philosophers, and argue in this paper using the methods of analysis and critical argumentation, that Socrates' arguments are not sufficient to justify obedience.

Keywords: Socrates, Arguments, Political Obligation, Reconsideration.

1. Introduction

The actual and concrete human society is basically divided into two spheres. The first is the sphere of the government or the state, which serves as controlling unit; the second, is the sphere of the governed, comprising the citizens or individual human beings under the control of the government. The main relationship between the two can be reduced to the relation between law and obedience, which in the proper sense, can be explained in terms of political obligation or duties of the citizens to the government and vice versa – the government makes laws; the citizens are expected to obey the law. Defining this relationship is a significant problem in political philosophy. It is a fundamental problem because there are normative, moral and political frictions resulting from the state-citizen social intercourse, especially on this very issue of political obligation.

*Department of Philosophy, Obafemi Awolowo University, Ile-Ife, 220005; okoriendu@yahoo.co.uk; ORCID ID: http://orcid.org/0000-0003-3098-3303

**Department of Philosophy and Religious Studies, Augustine University, P.M.B. 1010, Ilara-Epe 106101; abidemi.ogunyomi@augustineuniversity.edu.ng; ORCID ID: https://orcid.org/0000-0002-1911-3334

Alan Gewirth categorizes the basic questions of political philosophy into three. The first is a general question about society: why should human beings live in society at all? The second is a general question about government: why should human beings obey any government at all: why should some human beings have political power over others? The third is a specific question about government (and it is divided into three parts): what are the criteria for determining who should have political power? What should be the criteria for determining the extent or limit of political power? What ends do political power aim at and what are the criteria for determining those ends (Gewirth, 1965, p. 4)? The second question about obedience or political obligation is our interest in this paper. The reason is that this question has been reoccurring persistently from ancient to contemporary times. For instance, the question was originally raised during the Socratic cum Platonic era and discussed by other philosophers from the ancient, through the modern to the contemporary period, such as: Aristotle, Augustine, Thomas Hobbes, J. J. Rousseau, John Locke, J. S. Mill, and Immanuel Kant. However, while these philosophers have addressed it from different perspectives, the problem of political obligation remains an open-ended issue about which the final words have not been written nor mentioned in political philosophy. We therefore retrace the problem to ancient times when it was originally raised, and attempt to critically reexamine the arguments canvassed for it.

Are citizens morally obliged to obey the laws? Or put differently, what rationalizes political obligation? Why should citizens obey the law? This question is yet to be settled definitively. Dagger and Lefkowitz (2014) contend that to have a political obligation is to have a moral duty to obey the laws of one's country or state and on that point, there is almost complete agreement among political philosophers. Dagger and Lefkowitz may be right in saying there is almost complete agreement among political philosophers on this matter. However, political philosophers do not agree on the principle that rationalizes or justifies political obligation.

In the contemporary period however, the debate has taken various forms: between libertarianism and communitarianism, individualism and paternalism, absolutism and perfectionism, and so on. The central problem of the debate, nonetheless, is the same problem of political obligation. As a result of this, there are numerous conceptions and theories of political obligation. Our main concern in this paper is to explore the historical relevance of the problem of political obligation and critically re-examine Socrates' conception of the problem, stating why his arguments for political obligation do not justify political obligation.

The paper is structured into four major sections excluding the introductory and the concluding sections. The first section is a background to Socrates' conception of the problem of political obligation. This is important as this background will bring out the nature of the problem as originally raised, seeing how Socrates defines this perennial problem in political philosophy. The next section will be on the historical significance of the problem of political obligation. This section will explain how the problem first arose in the history of philosophy. The next section will attempt to identify and present Socrates' arguments for political obligation and cross examine his arguments with the views and responses of other political philosophers in various epochs of philosophy that succeeded the ancient period to some of the cogent issues raised in Socrates' arguments. The next section will carry out the

task of re-examining and evaluating the arguments of Socrates. This section shall canvass arguments to establish the thesis that Socrates' arguments are insufficient to justify political obligation. After this is done, a brief summary of the major issues undertaken in the paper will be outlined, which will serve as the conclusion.

2. Background to Socrates' Conception of Political Obligation

In Plato's *Apology*, which dwells mainly on Socrates' defense in Athenian court, Socrates defends himself against the accusations of Anytus and Meletus. He divides the accusations into two parts: the older charges and the later ones. The older charges against Socrates, according to Plato, are that:

> he is an evil-doer, and a curious person, who searches into things under the earth and in heaven, and he makes the worse appear the better cause; and he teaches the aforesaid doctrines to others charging them for the teachings (Plato (380 BC, p. 3).

In other words, Socrates' older accusers accused him of being overambitious, dubious, and manipulative. These, for Socrates, are widespread and age-old accusations, believed by almost everyone in Athens from childhood onward. He contends that the accusers are many, but the only person he remembers among them is a comic poet, called Aristophanes (Plato, 380 BC, p. 3). Socrates mentions that he is more afraid of the older accusers than the later ones. To put it in his words:

> These are the accusers whom I dread; for they are the circulators of this rumor, and their hearers are too apt to fancy that speculators of this sort do not believe in the gods. And they are many, and their charges against me are of ancient date, and they made them in days when you were impressible – in childhood, or perhaps in youth – and the cause when heard went by default, for there was none to answer. And, hardest of all, their names I do not know and cannot tell; unless in the chance of a comic poet (Plato, 380 BC, p. 3).

Socrates confesses that he is more afraid of the first set of accusers because in the first place, their hearers have a general notion that anyone charged with such accusation does not believe in the gods; second, they are so many to the extent that he does not know them; third, they have inculcated a bad impression about him in the Athenians right from their childhood, through their false accusations; fourth (and most pathetic for Socrates), they are not available for interrogation in the court of law.

He therefore defends himself against the first charges, stating that he does not venture into anything different from natural philosophy and he is not like the sophists who charge people for their teachings, while the people who have listened to his teachings before bear witness to his innocence. Socrates explains before the court that all those erroneous impressions about him are as a result of Chaerophon's consultation at the temple of a god called *Delphi* as to whether there is anyone wiser than he (Socrates) in Athens, which the god denies. He therefore takes it up as an assignment to prove the god wrong by engaging the wise people. In his attempt to find out their professed wisdom, he discovers their ignorance,

which makes him think that the god calls him the wisest because he himself is aware of his own ignorance while they are not.

The later accusers are: Meletus, Anytus and Lycon. And the accusations are: that Socrates is a doer of evil, corrupter of the youth, a disbeliever in the gods of the State, possessor of other new personal gods (Plato, 380 BC, p. 8). Socrates therefore, challenges Meletus, who is the leader of the accusers to tell him who the improver of the youth is if he is the only one that corrupts them. He asks Meletus to:

> Tell the judges, then, who is their improver; for you must know, as you have taken the pains to discover their corrupter, and are citing and accusing me before them. Speak, then, and tell the judges who their improver is. Observe, Meletus, that you are silent, and have nothing to say. But is not this rather disgraceful, and a very considerable proof of what I was saying, that you have no interest in the matter? Speak up, friend, and tell us who their improver is (Plato, 380 BC, p. 3).

Meletus is not able to tell the court precisely who the improver of the youth is. And that makes him ridiculous before the court. Socrates asks further questions, which leads Meletus into self-contradiction. The questions are:

> Which is better, to live among bad citizens, or among good ones? Answer, friend, I say; for that is a question which may be easily answered. Do not the good do their neighbors good, and the bad do them evil? Certainly. And is there anyone who would rather be injured than benefited by those who live with him? Answer, my good friend; the law requires you to answer – does anyone like to be injured? Certainly not. And when you accuse me of corrupting and deteriorating the youth, do you allege that I corrupt them intentionally or unintentionally? Intentionally, I say. But you have just admitted that the good do their neighbors good, and the evil do them evil. Now is that a truth which your superior wisdom has recognized thus early in life, and am I, at my age, in such darkness and ignorance as not to know that if a man with whom I have to live is corrupted by me, I am very likely to be harmed by him, and yet I corrupt him, and intentionally, too? (Plato, 380 BC, p. 9).

Socrates' defense in the above is that human beings prefer to live among good citizens, who will do them good, to bad ones who will do them evil. If that is the case, then no citizen will intentionally corrupt the people among whom he lives, knowing that if they are corrupt, then they will do him harm and not good. The purpose of this is to show that there is self-contradiction in the accusation leveled against him as the corrupter of youth.

He defends himself against the second charge as an atheist, and argues that Meletus attributes Anaxagoras' statement that the sun is stone and the moon is earth to him, thereby fooling the judged himself (Plato, 380 BC, p. 10). He professes his belief in the gods before the court and his commitment to the responsibility they have assigned to him, namely his philosophical quests. He therefore tells the court that even if the court is willing to set him free on the condition that he gives up his philosophical quests, he will obey the gods, rather

than the court (Plato, 380 BC, p. 13). The court finds him guilty after casting a vote and gives him the honor of proposing his own punishment.

However, because of his strong conviction that he has not offended the state, he proposes to be treated with dignity and honor in the Prytaneum like eminent Athenians, generals, Olympian winners and other outstanding people (Stumpf & Fiester, 2003, p. 45). He argues:

> There can be no more fitting reward than maintenance in the Prytaneum, O men of Athens, a reward which he deserves far more than the citizen who has won the prize at Olympia in the horse or chariot race, whether the chariots were drawn by two horses or by many. For I am in want, and he has enough; and he only gives you the appearance of happiness, and I give you the reality. And if I am to estimate the penalty justly, I say that maintenance in the Prytaneum is the just return (Plato, 380 BC, p. 19).

Socrates' proposed honorable treatment for himself instead of punishment infuriates the judge, and he condemns him to death. This is a practical example of friction or clash between the interest of an individual and the interest of the state. Individuals in society have different tasks, aspirations and focuses. In a situation where the state frowns or legislates against the aspiration of an individual, moral and normative issues about compliance or revolt ensue. In the case of Socrates, he has a divine assignment to carry out his philosophical quests. The state, on the other hand, frowns on such an assignment. Where, then, do we draw the line? This is the core issue in political obligation. Now, we move to the next section where we shall briefly articulate the historical account of how philosophical discussion of the problem of political obligation first arose in the history of philosophy.

3. Historical Significance of the Problem of Political Obligation

Political obligation is one of the major problems in contemporary political philosophy. According to Hopton (1998, 596), some philosophers like Thomas Hill Green have argued that political obligation is the central problem of political philosophy. It, however, has a very long historical origin dating far back to the Socratic era in ancient Greece. Socrates makes a creative redirection of the philosophical engagement from the speculative endeavors of his predecessors to the practical issues in human society. A part of those practical issues to which Socrates directs the attention of philosophy is the problem of political obligation. Surprisingly, he is the first "specimen" in the experimentation of the discourse. Political obligation, by definition, refers to a moral duty or requirement to obey the state or national laws. It is a moral duty of the citizens to obey the laws of their land.

In the *Crito*, Socrates refuses to escape from prison against the strong advice of his friend, Crito, even when he has the chance to bribe the warders and escape, after being unjustly sentenced to death by Athenian law. However, Crito's persistent effort to convince him to escape provokes a normative discussion between Socrates and Crito on the implications of such action. The discussion results in the rationalization of political obligation by Socrates.

Before Socrates, it is doubtful whether there was any such strict justification of political obligation. Hence, the problem of political obligation has a significant historical relevance and deep root in Socratic philosophy.

4. Socrates' Arguments for Political Obligation

Despite the fact that the Athenian laws condemn Socrates to death unjustly, obedience to the law is still reasonable to him. In his attempt to rationalize political obligation, Socrates puts some arguments forward. These arguments are supposed questions which the laws of Athens could ask him if he decided to follow Crito's injunction to escape from prison.

For us to present a clear explanation of the Socratic conception of political obligation, we shall dwell on Plato's *Crito*. We shall also quote heavily from it. Crito expresses anxiety over what people will say if he is not able to save Socrates from the death penalty. He says they may call him a stingy fellow that places much value on money rather than the life of his friend, not knowing that it is Socrates who refuses to escape on his advice. This marks the beginning of serious arguments between Socrates and Crito about the moral or immoral status of such action.

Socrates' arguments are predicated upon the following principles: that the opinion of some men and not the majority ought to be regarded. For him, there is no good opinion with the majority: they are incapable of hurting a man or making him good. The opinion of the wise is good while the opinion of the unwise is evil. The opinion of the wise ought to be followed when confronted with such diametrically opposed concepts as right and wrong, just and unjust, good or bad, honorable and dishonorable, etc.

The second principle is that not just mere life but a good life matters. And a good life, for Socrates, is equivalent to a just and honorable life. So a just and honorable life ought to be valued even more than life itself.

The third principle is that it is wrong to do wrong intentionally. When injured, it is not right to retaliate. Retaliating injustice for injustice or evil for evil is wrong. Socrates' first argument for political obligation, as recorded by Plato (2007, 6–7), therefore, goes thus:

> ... If the laws and the community of the city came to us when we were about to run away from here ... and standing over us were to ask, "Tell me, Socrates, what are you intending to do? By attempting this deed, aren't you planning to do nothing other than destroy us, the laws, and the civic community, as much as you can? Or does it seem possible to you that any city where the verdicts reached have no force but are made powerless and corrupted by private citizens could continue to exist and not be in ruins?"

From the above expressions, Socrates emphasizes the moral supremacy of the community over the individual citizens living in it. No community will survive if its laws are consistently violated and broken by the individuals in it. And no individual should disobey the law just for the reason that the law violates his personal interest; for if every individual in a society engages in the same act, society will crumble and degenerate into a state of anarchy. Any act of revolt from the individuals in the society against the laws translates to bringing the society itself into ruin or destruction. Whether or not the law is just, whatever the nature

of the law, individuals must obey. Socrates takes a radical communitarian stance in this argument, which both places the interest of the state above the interests of the individuals living in it and asserts that even unjust laws ought to be obeyed.

The above view has, however, been criticized by medieval philosophers like Thomas Aquinas and classical utilitarians like J. S. Mill. For Aquinas, there are two types of law: just and unjust laws. Unjust laws are acts of violence rather than laws and they ought not to be obeyed. In fact, like Augustine, Aquinas (1911, p. 1365), contends that an unjust law is no law at all. J. S. Mill, on the other hand, contends there are two spheres of morality: the private and the public. The private belongs to an individual while the public belongs to the state. The state ought not to interfere with the affairs of an individual that affects him/her directly because if it does, it will be wrong. The state can only persuade an individual to behave in certain ways but should not compel an individual to do so. The only thing that warrants the legislation of the state against the actions of an individual is when such actions portend harm to others (Mill, 1971, pp. 270–274; Midgley, 1991, pp. 31–38). J. S. Mill, however, agrees with Socrates that individuals who have enjoyed some benefits from the state owe a debt of gratitude to the state. However, this gratitude may not necessarily manifest as obedience to the law (Mill, 2001, pp. 69-71). Apart from that, contemporary philosophers like John Rawls have argued that everyone possesses an inviolability founded on justice that the interest of the state as a whole cannot override (Rawls, 1999, p. 3). Socrates' radical communitarian perspective on political obligation sacrifices the interest of the individual for that of the state. Socrates' second argument for political obligation goes thus:

> . . . What if the laws then said, ". . . what reason can you give us and the city for trying to destroy us? Did we not, to begin with, give birth to you? And wasn't it through us that your father married your mother and conceived you? So show those of us, the laws concerning marriages, what fault you find that keeps them from being good?" "I find no fault with them," I would say. "What about the laws concerning the upbringing and education of children, by which you too were raised? Or didn't those of us, the laws established on this matter, give good instructions when they directed your father to educate you in the arts and gymnastics?" . . . And if this is so, do you suppose that justice between you and us is based on equality, and do you think that whatever we might try to do to you, it is just for you to do these things to us in return? . . . if we try to destroy you, thinking this to be just, you will then try to destroy us the laws and your homeland in return. . . ? Are you so wise that it has slipped your mind that the homeland is deserving of more honor and reverence and worship than your mother and father and all of your other ancestors? And is held in higher esteem both by the gods and by men of good sense? . . . And that whether it is to be beaten or imprisoned, or to be wounded or killed if she leads you into war, you must, do it? And that . . . in the courts and everywhere you must do what the city and the homeland orders. . . ? And that it is not holy to use force against one's mother or father, and it is so much worse to do so against one's homeland?" What will we say to this, Crito . . . (Plato, 2007, pp. 7–8)?

Socrates' second argument for political obligation above dwells basically on the question of benefit. An individual ought to obey the law as a sign of gratitude to the state for all the benefits, such as security, education, and marriage – which the state has bestowed upon him from childhood to adulthood. The only way to compensate the state for its kindness is to respect and obey its laws, given the benefits individuals receive from infancy. Disobedience to the laws amounts to ingratitude and injustice. If one disobeys the laws after enjoying its benefits, one will be acting against its interests, and it is ingratitude to act against the interests of one's benefactor.

In addition, the passage emphasizes the paternalistic status of the state to the individuals in it and asserts that nothing at all justifies disobedience to the laws of the state. It is improper to revolt against the laws in the name of fighting for equality because, from Socrates' point of view, an individual cannot compete side-by-side for equality with the laws. Socrates places more value on obedience to the law than obedience to the instructions of one's parents.

In a situation where there is a clash between the instructions of one's parents and the instructions of the state, priority ought to be given to the instructions of the state. One ought to do whatever the laws of one's state requires one to do without hesitation, even if such obedience will lead to death. Socrates' justification of political obligation here is based on benefit or gratitude and the paternalistic status of the state.

This argument has been extensively discussed and advanced by Smith (1973), Walker (1988) and Klosko (1989). M. B. Smith, for example, argues, first, that most benefits that the state confers on its citizens are intended to achieve aims other than the interest of its citizens. Apart from that, the state does not consider whether or not its citizens really want those benefits. For these reasons, he argues that he is not under any obligation to be grateful for enjoying benefits which aim at achieving interests other than his welfare (Smith, 1973, p. 953).

Second, he argues, even if citizens owe an obligation of gratitude to the state, such gratitude is not a genuine one. What makes genuine gratitude indisputable for Smith is cheerfulness and willingness (Smith, 1973, p. 953). But a person owing a debt of gratitude cannot be said to express it cheerfully or willingly, because paying a debt which one owes another person is compulsory whether or not one likes it. For this reason, if a citizen's obedience to the laws of a state does not proceed out of willingness and happiness, it cannot correctly be called gratitude.

While A. D. M. Walker justifies the gratitude theory, Gregory Klosko sees Walker's arguments as unconvincing and the entire project of gratitude theory as weak, though real. For Walker, if citizens are beneficiaries of the state, they ought to be grateful to the state. And in being grateful, they ought not to act contrary to the interests of the state (Walker, 1988, p. 205). But what exactly does Walker mean by the phrase: "interests of the state"? He does not seem to have set any boundary as to what such interest could be and the reality of an individual succumbing to such interests that has no restriction.

George Klosko is against this view. For Klosko, the problem with Walker's gratitude theory is that though it exists, it is generally weak and diffuse. It can be overridden frequently

and as such, it cannot serve as a genuine theory of political obligation. It does not require compliance with onerous and burdensome laws (Klosko, 1989, pp. 354–355). In other words, in the face of complex laws, gratitude theory is not to be taken seriously. An example of such an onerous and burdensome law is the issue of taking one's life unjustly, which has been our concern in the previous discussion. Let us consider the third argument:

> ... and yet even so we pronounce that we have given the power to any Athenian who wishes, when he has been admitted as an adult and sees the affairs of the city and us the laws and is not pleased with us, to take his possessions and leave for wherever he wants. And if any among you wants to live in a colony because we and the city do not satisfy him, or if he wants to go somewhere else and live as a foreigner, none of us laws stand in the way or forbids him from taking his possessions with him and leaving for wherever he wants. But whoever remains with us, having observed how we decide lawsuits and take care of other civic matters, we claim that this man by his action has now made an agreement with us to do what we command him to do, and we claim that anyone who does not obey is guilty three times over, because he disobeys us who gave birth to him, and who raised him, and because, despite agreeing to be subject to us, he does not obey us ... (Plato, 2007, p. 8).

Here, Socrates' argument for political obligation is based on the notion of contract or consent. If an individual was born and nurtured in a particular society and this individual, after coming of age, lives, gets married, procreates, and establishes himself in that same society, despite having the option to migrate to another society, change his citizenship or go on exile, this person, in effect, has agreed, consented or covenanted (at least implicitly) to abide by whatever rules and regulations govern the existence of this society. So, any individual that disobeys the laws of the state violates the covenant he has freely made with it. From a Socratic point of view, this amounts to injustice against the state. Therefore, political disobedience is unjustified.

Modern philosophers like Thomas Hobbes and John Locke have developed this consent theory, using the hypothesis of the state of nature. While Thomas Hobbes' political obligation is necessitated by the fear of violent death in the nasty and brutish condition of the state of nature and the urgent need to seek peace by entering into an irrevocable contract with the sovereign (Hobbes, 1651, pp. 78-81), John Locke's political obligation is necessitated by the need to secure life and property in the state of nature. Security of life and property, for Locke, is the main consideration for political obligation. This is the reason why individuals in the state of nature give their executive rights to govern themselves to the sovereign in human society. The contract is, however, revocable. Once the condition of security of life and property is defeated, political obligation is defeated (Locke, 1826, pp. 159–161). This implies that for Locke, the contract is not as absolute as Socrates' and Hobbes's contractarianism understand it to be.

J. J. Rousseau is another contractarian whose position is compelling. His introduction of the notion of "general will," which seems to make a provision for a unanimous opinion or decision of everyone regarding what is to be regarded as a binding or unbinding law, makes his account more appealing (Bertram, 2004, pp. 75–80). However, we have to note that a

more sophisticated explanation was provided by Immanuel Kant, who advocates a free and open discussion of what should count as legally binding and legally unbinding, that, in the final analysis, would eventuate in a revision and reformulation of laws, that would reflect the general consensus of the people (William, 1994, pp. 137–141; Kant, 1996, pp. 290–301).

Kant's position is predicated on the view that freedom, in the first place, is the basis of being a legitimate member of the society. This freedom grants one the right to make a public use of one's reason in all matters, and also places on one an obligation to obey the agreement, the making of which one is a part (Kant, 1784, pp. 2–6). Kant's version seems to be a modification of the contractarianism of his predecessors. It is, in particular, a careful rebrand of Rousseau's version. While his contractarianism avoids the difficulty of attributing absolute power to the state at the expense of the citizens in the contractarianism of Socrates and Hobbes, it avoids the attribution of the right to disobey the state to citizens when it is incapable of ensuring the security of lives and properties in the contractarianism of Locke. Again, by making freedom the basis of the contract, he transcends Rousseau's valorization of the supremacy of the general will, which, in Rousseau's view, justifies the compelling of citizens to obedience. For Rousseau, to compel the citizens to obey the general will is to force them to be free (Bertram, 2004, p. 87; pp. 194–195). However, for Kant, the freedom of the citizens is what brings the general will about in the first place. Their freedom to discuss and determine what should be binding and unbinding along with others, brings about the understanding that they have the obligation to respect the consensus they made with others as free beings and the understanding that the society cannot function well if such consensus is not respected by the citizens who are equal and free participants in its making.

5. Reconsidering Socrates' Arguments for Political Obligation

The first argument is for the supremacy of the community over the individual and the injunction to obey even unjust laws. This argument makes Socrates' exact position on the problem of political obligation unclear. In the *Apology*, Socrates argues that even if the law will set him free on the condition that he give up his philosophical quests, he will not comply. This is because that philosophical quest itself is a divine assignment given to him by the gods. And he prefers to obey the gods to the law. This is a strong act of political disobedience on Socrates' part.

However, in the *Crito*, Socrates argues that society is supreme over the individuals living in it. And its laws, whether just or unjust, ought to be obeyed. This, in light of contemporary trends in political philosophy, can be considered as a radical communitarian justification of political obligation. This makes it confusing whether to classify Socrates as an advocate of civil disobedience or political obligation. Nonetheless, William Lawhead attempts to harmonize Socrates' conflicting claims. According to him, Socrates, in his trial, made one exception to his principle of unconditional obedience to the law; arguing before the jury, he made it clear that he would refuse to obey any order to give up philosophizing, even if obedience would mean that he would be acquitted and disobedience would mean death. However, he saw this situation as different from illegally escaping from prison (Lawhead, 2002, p. 42).

His principal stance in the trial was that the gods commanded him to do philosophy and obedience to the gods took precedence over the state. Also, continuing to do philosophy would benefit the state by showing where it was wrong. Finally, continuing to philosophize would show respect for the law because, even though it would challenge the edicts of the court, its public nature would demonstrate a willingness to accept the legal consequences. By contrast, escaping from prison would be an attempt to evade the law by doing an end run around it (Lawhead, 2002, p. 42). That is a brilliant harmonization of conflicting claims. However, it does not dispel the fact that Socrates advocates civil disobedience and obedience to laws at the same time. According to Stephen Nathanson,

> Socrates' arguments in the *Crito* are often taken as a source of the view that people must always obey the law. Socrates appears to argue that people must always obey the law because the state is like a parent and one must obey one's parents, that the state has benefited him and therefore should be obeyed, and that he has made a tacit agreement to obey the laws by living in Athens all his life. In the *Apology*, however, Socrates states that he will disobey the law if it requires him to violate the commands of the gods. Socrates, then, is a source of both the individualist tradition that approves civil disobedience and the authoritarian, statist tradition that condemns it (Nathanson 2005, 260).

From the above, Socrates is seen as both the historical patriarch of civil disobedience and political obligation. This is because of the two conflicting positions he holds concerning the law in the *Crito* and in the *Apology*. We guess that is the reason a strong advocate of civil disobedience like Martin Luther King Jr. argues in his *Letter from Birmingham Jail* that "*. . . to a degree, academic freedom is a reality today because Socrates practiced civil disobedience*" (King, 2008, p. 194). We therefore reason along with both Stephen Nathanson and King that while Socrates is the patriarch of political obligation on the one hand, he is also the patriarch of civil disobedience on the other.

However, we may not justifiably credit Socrates with being the patriarch of civil disobedience. In the same way, we may not justifiably praise him for being the patriarch of political obligation or obedience, because, in fact, he likely was not aware of the two concepts in his time or that people would be influenced by his action(s) in the future. He was just a man who acted according to what he considered to be right, moral and virtuous. Apart from this, civil disobedience is not bad or detestable in itself. John Rawls sees it as a stabilizing device in a constitutional or democratic state (Rawls 1999, pp. 335–336). In other words, it is an instrument which the minority uses to checkmate the tyranny of the majority in a legitimate or well-ordered society. Before Rawls, Mahatma Gandhi and Martin Luther King Jr. adopted civil disobedience as a systematic method to address the excesses of the state. It is the last resort after other attempts have proved abortive in addressing the phenomenon of injustice in society (Nojeim, 2013, pp. 94–96; pp. 189–206).

Nonetheless, the condition upon which Socrates bases his civil disobedience appears to be unjustifiable. He prioritizes obedience to the divine elements, which are unverifiable in the empirical world over the laws of tangible human beings in society. This action contradicts his redirection of the philosophical enterprise from the superstitious and supernatural focuses of his predecessors to the practical issues in human society.

The second argument is from gratitude and the paternalistic status of the state. Gratitude theory may seem impressive as a justification for political obligation. However, we must consider whether losing one's life unjustly measures equally with the benefit which one has received from the state. In other words, it is not a balanced equation to lose one's life unjustly as an expression of gratitude to the state. Whatever the benefits of the state, one cannot reasonably consider losing one's life unjustly as a way of expressing gratitude to the state.

Furthermore, there is a need for us to distinguish between being morally obliged and being morally obligated. Being obliged is somehow subtle while being obligated is stringent. If one enjoyed certain benefits from the state or a friend, for instance, one might be said to be morally obliged to be grateful to the state or that friend. However, it does not follow that one is morally obligated. In this context, to be morally obliged is to consider oneself bound to give a courtesy requital to a benefactor. Failure to do so does not translate to injustice. However, to be morally obligated is to be legally bound to repay a benefactor or be grateful for the benefit enjoyed from the benefactor. Failure to do so amounts to injustice. Hence, to be morally obligated is to have a moral duty to perform. Immanuel Kant reminds us that to have a moral duty is to have the duty to act out of reverence for the law (Kant, 1964, pp. 20–21; pp. 67–68). In essence, one is morally obliged to be grateful to the state after one has enjoyed its benefits but one is not morally obligated to do so as Socrates argued.

The third argument is from the notion of contract. Socrates argues that one has entered into an implicit contract or agreement with the state if one has established themselves there when one has an option to go into exile or migrate to another society. This may hold. However, it holds as long as the state ensures security of life and property and leaves room for the achievement of the aspirations and self-fulfillment of the individuals living in it. The agreement is an implied one anyway, not a real or explicit one. Again, if citizens do not trust their government as a result of the inability of such government to protect lives and property, then, there is no justification for maintaining that citizens should obey the laws emanating from such government or state, because the "*convenient articles of peace upon which men may be drawn to agreement*," to use the parlance of Thomas Hobbes, have already been broken (Ogunyomi & Ogundele, 2021, p. 80). The allegiance and obedience owed to the state would not only be undermined, but would be completely defeated as a result of this inability of the state to secure life and property. This, in fact, should mark the end of the relationship between the state and citizens according to Lockean contractarianism. We therefore agree with Locke that political obligation is defeated once the state can no longer ensure the security of life and property.

Again, if the state does not serve the interests of the individuals but seeks instead to take their lives unjustly, that is the end of such an implied contract. An individual's decision to break with such a society does not amount to injustice. Self-interestedness is a peculiar attribute of human beings. It does not matter whether they inhabit the same society or not. Within the same society, all individuals, under the leadership of the state, have their various dreams, mindsets, aspirations, goals and focus, which may, in most cases, be entirely different from what the state as a whole calls the common good. While the state is committed to ensuring and securing the common good, it ought not to gloss over or

override the personal aspirations of individuals living in it; instead, it ought to provide an atmosphere where individuals in the society will feel fulfilled within the confines of the laws governing the society. This is consistent with John Rawls's submission that the individuals in a state possess an inviolability founded on justice, which the interest of the society as a whole cannot override (Rawls, 1999, p. 3).

6. Conclusion

In the foregoing, we have discussed political obligation, its background and its historical significance. We have discussed Socrates' three main arguments for political obligation: the argument from the supremacy of the state and the injunction to obey unjust laws, the argument from gratitude and the paternalistic status of the state, and the argument from implicit contract in light of the views of other political philosophers that succeeded Socrates across various epochs in the history of Western philosophy. We have attempted a critical reconsideration of the arguments of Socrates in support of obedience to the law. We argued that the exact position of Socrates on political obligation is not clear enough. This argument is predicated on the reason that he encourages civil disobedience in the *Apology* but encourages obedience to the law in the *Crito*. We have also argued that losing one's life unjustly is not an equal measure with whatever gratitude the state might have bestowed on one and that one may be morally obliged to be grateful to the state for its benefits, but one is not necessarily morally obligated to do so. We have also argued that even if there is an implied contract between the state and the individuals living in it, such a contract holds as long as the state maintains the security of life and property and sustains an intelligible avenue where the individuals' aspirations, aims and genuine life pursuits can be promoted without upsetting the common good.

References

Aquinas, T. (1911). *Summa Theologica*. http://www.ccel.org/ccel/aquinas/summa.html

Bertram, C. (2004). *Routledge philosophy guidebook to Rousseau and the social contract*. Routledge.

Dagger, R., & Lefkowitz, D. (2014). Political obligation. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Fall 2014 ed.). Stanford University. https://plato.stanford.edu/entries/political-obligation/

Gewirth, A. (1965). *Political philosophy*. Macmillan.

Hobbes, T. (1651). *Leviathan, or, the matter, form, & power of a commonwealth ecclesiastical and civil*. Printed for Andrew Crooke.

Hopton, T. (1998). Political obligation. In R. Chadwick (Ed.), *Encyclopedia of applied ethics* (Vol. 3). Academic Press.

Kant, I. (1784). *An answer to the question: What is enlightenment?* Blackmask Online. http://www.blackmask.com

Kant, I. (1964). *Groundwork of the metaphysic of morals* (H. J. Paton, Trans. & analysis). Harper & Row.

Kant, I. (1996). *Practical philosophy* (M. J. Gregor, Ed. & Trans.). Cambridge University Press.

King, M. L., Jr. (2008). Letter from Birmingham jail. In C. Carson (Ed.), *The autobiography of Martin Luther King, Jr.*

Klosko, G. (1989). Political obligation and gratitude. *Philosophy & Public Affairs, 18*(4), 352–358.

Lawhead, F. W. (2002). Socrates. In *The voyage of discovery: A historical introduction to philosophy* (2nd ed.). Wadsworth.

Locke, J. (1826). *Two treatises of government*. Printed for Thomas Tegg; W. Sharpe and Son.

Midgley, M. (1991). *Can't we make moral judgements?* Bristol Press.

Mill, J. S. (1971). The individual and the limits of state power. In S. E. Stumpf (Ed.), *Philosophical problems: Selected readings* (3rd ed.). McGraw-Hill.

Mill, J. S. (2001). *On liberty.* Batoche Books.

Nathanson, S. (2005). Civil disobedience. In D. M. Borchert (Ed.), *The encyclopedia of philosophy* (2nd ed., Vol. 2). Macmillan Reference.

Nojeim, M. J. (2013). *Life and works of Gandhi and Martin Luther King Jr.: The power of nonviolent resistance.* Beulahland Publications.

Ogunyomi, A. I., & Ogundele, E. A. (2021). Aristotle, Confucius, and Rousseau on human nature and the golden mean: A comparative analysis. *Prajna Vihara: Journal of Philosophy and Religion, 22*(1), 71–84.

Plato. (380 BCE). *The apology of Socrates* (B. Jowett, Trans.).

Plato. (2007). *The Crito* (C. Woods & R. Pack, Trans.). Creative Commons.

Rawls, J. (1999). *A theory of justice* (Rev. ed.). Harvard University Press.

Smith, M. B. (1973). Is there a prima facie obligation to obey the law? *Yale Law Journal, 82*(1), 950–976.

Stumpf, E. S., & Fiester, J. (2003). *Socrates to Sartre and beyond: A history of philosophy* (7th ed.). Kenneth King.

Walker, A. D. M. (1988). Political obligation and the argument from gratitude. *Philosophy & Public Affairs, 17*(3), 191–211.

Williams, H. (1994). Kant on the social contract. In D. Boucher & P. Kelly (Eds.), *The social contract from Hobbes to Rawls* (pp. 135–158). Routledge.

Balkan Journal of Philosophy (2025), vol. 17, issue 2, 228–232

TOCQUEVILLE'S DEMOCRACY: THE PROMISE AND PERILS OF EQUALITY

David Pimenta*

Abstract: This essay offers a critical engagement with *Democracy in America*, examining the enduring relevance of Alexis de Tocqueville's reflections on equality and its intricate relationship with liberty. By reconstructing his analysis of the transition from aristocracy to democracy, it highlights the ambivalent nature of equality, as both a force for social advancement and a latent threat to individual freedom. The essay situates these insights within the broader structural and psychological tensions of modern democratic regimes, where the pursuit of autonomy can paradoxically foster dependence on centralized authority and weaken civic engagement. Addressing key concepts such as majority rule, soft despotism, and the moral foundations of political participation, the essay underscores the necessity of cultivating institutional, cultural, and ethical resources capable of sustaining liberty under the pressures of egalitarianism.

Keywords: Alexis de Tocqueville; Democracy in America; Democracy; Equality; Liberty.

Amid the growing strains on liberal democracy worldwide – namely, rising political polarization and institutional erosion – Alexis de Tocqueville's *Democracy in America* offers a remarkably prescient account of the promises and perils inherent in modern egalitarian societies. This essay seeks to engage Tocqueville not merely as a historical observer of early American democracy but as a foundational thinker (Craiutu 2014), whose reflections on equality, liberty, and despotism continue to illuminate the structural tensions of democratic life. By reconstructing his analysis of the democratic condition, the essay aims to show why Tocqueville's political philosophy remains vital to contemporary debates.

Tocqueville was born in 1805, in the aftermath of the great republican revolutions – the American and the French – that heralded the end of a world dominated by monarchies and aristocracies. Educated in law, Tocqueville immersed himself in the study of classical texts and influential thinkers such as Machiavelli and Montesquieu, while attending the lectures of François Guizot (Craiutu 1999). This intellectual foundation endowed him with a profound historical sensibility, allowing him to anticipate the emergence of the democratic era and to discern its inherent challenges. Upon completing his studies, Tocqueville, accompanied by his friend Gustave de Beaumont, embarked in 1831 on a journey to the little-known *New World* with the explicit aim of examining the only extant democratic and liberal nation – America.

Following his journey, Tocqueville published his magnum opus, *Democracy in America*, amidst a transitional period from the world of the *Ancien Régime* to the democratic age. In this work, the author offers a nonpartisan analysis of the distinctions between the aristocratic

*Institute of Social Sciences, University of Lisbon, Portugal; davidpimenta@edu.ulisboa.pt; ORCID ID: https://orcid.org/0000-0001-8520-3058

world, destined for extinction, and the inexorable progression toward a democratic society. He advocates for the establishment of a new political science tailored to the demands of the emerging era.

The method employed by Tocqueville, inherited from Montesquieu, is comparative in nature, juxtaposing the two worlds of aristocracy and democracy. Like Machiavelli, Tocqueville examines politics as it is, rather than as it ought to be – a marked departure from the *Philosophes*, who were often inebriated by utopian ideologies. In this regard, Tocqueville refrains from defining democracy in abstract terms; instead, he describes it as a general and irreversible trend that manifests as a concrete phenomenon. This phenomenon is driven by a dynamic process of equalization of conditions, tending toward uniformity in lifestyles and standards of living, with the ultimate goal of achieving social well-being for the greatest number of individuals (Tocqueville 2002, 15–6).

Equality, therefore, emerges as the foundational principle of democracy, influencing all aspects of society and shaping the character of democratic individuals (Tocqueville 2002, 573). However, it is essential to note that this equality is neither absolute (Tocqueville 2002, 607) nor does democracy constitute a definitive endpoint. Unlike his contemporaries, such as Karl Marx, Tocqueville does not adopt a historicist perspective. Recognizing that democracy is rooted in equality – just as aristocracy was grounded in inequality – Tocqueville asserts that, while the democratic revolution is unstoppable, it remains the responsibility of human communities to embrace democracy with varying degrees of liberty. This reveals the Tocquevillian paradox: equality is compatible both with freedom and with serfdom, depending on how it is implemented and manifested within society (Tocqueville 2002, 785).

Tocqueville regards America as a model to be emulated by Christian nations transitioning toward democracy. In the *New World*, he observes firsthand that humanity can achieve a balance between equality and liberty[1].

Equality, much like democracy and aristocracy, possesses both positive and negative aspects, serving as the pivotal element in Tocqueville's argument. Its principal virtue lies in the democratic individual's aspiration to improve their condition, driven by the belief that they possess equal opportunities to attain the same status as their peers (Tocqueville 2002, 72). The negative aspect of equality arises from passions such as envy and selfishness[2], which can lead to a leveling down of societal standards of equality (Tocqueville 2002, 72).

The democratic regime, while presupposing equality, also entails individualism, wherein citizens assert their independence and withdraw from society through deliberate choice (Tocqueville 2002, 574); the outcome of this individualism is the atomization of society (Tocqueville 2002, 575). Consequently, the interdependence that characterized class relationships

[1]However, Tocqueville does not advocate for a blueprint of the American model. Beyond emphasizing the central role of human agency in the progression of the democratic revolution, he argues that each nation follows a distinct path of democratic development and that the social and political consequences of democratization vary from one country to another.

[2]A parallel argument to that of Rousseau can be observed here. The French philosopher depicts humanity as naturally free and good, but through a historical accident, humans transition from a nomadic to a sedentary existence. This shift leads to the development of social relations, which eventually give rise to passions such as admiration, selfishness, and revenge, all stemming from the phenomenon of comparison (Rousseau 2018).

during the aristocratic era, transforms, in the democratic age, into a dependence between each individual and the State. Tocqueville's diagnosis anticipates Francis Fukuyama's contemporary concern that, as intermediary structures erode, individuals increasingly seek identity and recognition not through civic engagement but through direct appeals to the state, fueling polarization and undermining liberal democratic cohesion (Fukuyama 2018).

This shift gives rise to a central problem: as individuals focus exclusively on the pursuit of their personal well-being, they enable the State to assume total power over the management of all remaining public and private affairs. Within this framework, Tocqueville identifies two fundamental issues inherent in a democratic regime: the tyranny of the majority and the emergence of the new despotism.

The tyranny of the majority manifests as a form of dictatorship, where the majority imposes its will on the minority within society. This majority exerts its influence not only in the legislative realm but also in shaping public opinion, defining what is considered good or bad. This dynamic cultivates an egalitarian, yet tyrannical culture that entraps the very souls of individuals (Tocqueville 2002, 575). The tyranny of the majority is not merely a political phenomenon addressable through political solutions; it is also a deeply rooted social reality. Social facts, particularly public opinion, are the most challenging to constrain, considering that opposing the collective judgment of society is an arduous task.

The new despotism, in turn, is a "soft" despotism, willingly accepted by sovereign citizens who permit the unchecked growth of a colossal State that intervenes in all aspects of public life. This form of despotism resembles a tyrannical parent who assumes responsibility for everything, ensuring comfort and security but reducing citizens to a state of perpetual childhood (Tocqueville 2002, 770; Mansfield 2010).

Tocqueville proposes remedies to counteract the despotism inherent in democratic regimes by harmonizing liberty with equality. His solutions are multifaceted, encompassing mechanisms tied to individual interest, religion, and the active participation of civil society.

In America, Tocqueville observes the doctrine of self-interest rightly understood as spontaneously embraced by citizens. This doctrine asserts that each individual learns to reconcile their personal interests with those of the community. In other words, the individual acts rationally and moderately, pursuing personal goals in a measured way that ultimately benefits the collective. In this framework, virtues are esteemed as virtues precisely because they are useful (Tocqueville 2002, 594).

As for religion, while it is not inherently a political element, it possesses the capacity to influence political life; in fact, it is closely tied with liberty (Mansfield 2016). A strong and independent religion provides individuals with morality and convictions, thereby tempering and guiding democratic societies.

Regarding the moderating elements of civil society, several key factors stand out: the right to free association, which protects minority rights and transcends civic mediocrity; the freedom of the press; local self-government, which fosters citizen accountability and reduces state centralization; the independence of the judiciary, and the separation of powers maintained through a system of checks and balances.

Tocqueville further develops his argument by warning about the two tendencies arising from equality (Tocqueville 2002, 743).

The first tendency is anarchic in nature, characterized by the absence of any individual or entity deemed legitimate to wield authority. It is revolutionary, as despotism emerges through the erosion of the rule of law. However, this tendency can be effectively restrained by the aforementioned remedies provided by the art of politics.

The second tendency represents the primary threat and unfolds insidiously. The democratic individual assumes that all power granted to the government ultimately belongs to them as a sovereign being. This belief leads the individual, in their insatiable pursuit of personal well-being, to relinquish increasingly larger portions of their liberty to the government. As a result, the State becomes a more powerful entity, surpassing any historical despotism exercised by monarchs or aristocrats, as it assumes ever-expanding functions.

The same tendency that prevents the democratic individual from acknowledging any authority beyond themselves simultaneously compels them to assign unlimited authority to the central government, which acts in the name of collective equality. It is thus essential that the democratic citizen develop a reflective sense of responsibility, mindful that the very force which promises liberation may, in its excess or misdirection, become a source of constraint. When despotism is concentrated in a single ruler or governing class, citizens can identify and confront the enemy. However, when the despot is an immense, faceless entity like the State, reforming the prevailing regime becomes a far more daunting task. As the French poet and contemporary of Tocqueville, Alfred de Vigny, aptly observed: "The least bad government is the one that appears the least, is felt the least, and costs the least"[3] (Freitas do Amaral 1997).

Tocqueville's reflection on democracy is neither a nostalgic defense of aristocracy nor an uncritical embrace of modern egalitarianism. Rather, it is a philosophical inquiry into the paradoxes of equality: a force that simultaneously emancipates and entraps, that undergirds liberty yet risks inviting servitude. In identifying equality as the telos of modern political history, the author offers a theory of democratic condition grounded not in utopian abstraction but in the empirical observation of its tensions and ambivalences.

In a time when liberal democracies confront increasing internal and external pressures, this essay has foregrounded the dangers of democratic despotism, yet it also seeks to reclaim Tocqueville's constructive ambition: the development of a political framework capable of addressing the unique challenges posed by democratic equality. In doing so, it affirms the need for institutions, civic practices, and moral dispositions that counterbalance the leveling force of equality. The responsibility for maintaining freedom in the democratic era, Tocqueville reminds us, rests not only with the state but with its citizens.

Acknowledgements

David Pimenta acknowledges the financial support provided by the Portuguese Foundation for Science and Technology and the EU for his doctoral grant (reference UI/BD/153726/2022).

[3]English translation of "(L)e moins mauvais gouvernement est celui qui se montre le moins, que l'on sent le moins et que l'on paie le moins cher" (Vigny 2025).

References

Craiutu, A. (1999). Tocqueville and the political thought of the French doctrinaires (Guizot, Royer-Collard, Rémusat). *History of Political Thought, 20*(3), 456–493.

Craiutu, A. (2025). Tocqueville's new science of politics revisited: A few lessons for contemporary political scientists. *Online Library of Liberty.* https://oll.libertyfund.org/publications/liberty-matters/aurelian-craiutu-tocqueville-s-new-science-of-politics

Freitas do Amaral, D. (1997). *História das ideias políticas* (Vol. 2).

Fukuyama, F. (2018). *Identity: Contemporary identity politics and the struggle for recognition.* Profile Books.

Mansfield, H. C. (2010). *Tocqueville*: A very short introduction. Oxford University Press.

Mansfield, H. C. (2016). Tocqueville on religion and liberty. *American Political Thought, 5*(3), 250–276. https://doi.org/10.1086/685889

Rousseau, J.-J. (2018). *The social contract and other later political writings* (V. Gourevitch, Ed.). Cambridge University Press.

Strauss, L., & Cropsey, J. (Eds.). (1987). *History of political philosophy* (3rd ed.). University of Chicago Press.

Tocqueville, A. de. (2002). *Democracy in America* (Vols. 1–2; H. Reeve, Trans.). Pennsylvania State University Press. http://seas3.elte.hu/coursematerial/LojkoMiklos/Alexis-de-Tocqueville-Democracy-in-America.pdf

Vigny, A. de. (n.d.). *Le moins mauvais gouvernement est celui qui se montre le moins, que l'on sent le moins et que l'on paie le moins cher.* Citations Ouest-France. https://citations.ouest-france.fr/citation-alfred-de-vigny/moins-mauvais-gouvernement-celui-montre-22172.html

Balkan Journal of Philosophy (2025), vol. 17, issue 2, 233–248

DECODING THE POLITICAL IN LEVINAS AS OTHERWISE THAN POLITICAL

Anto Varghese*

Sreekumar Nellickappilly**

Abstract: This paper investigates Emmanuel Levinas' nuanced engagement with politics, highlighting his critique of Western philosophy's tendency toward totalizing frameworks. Levinas views the political failures of his time as rooted in these philosophical shortcomings, advocating for an "ethical turn" that departs from ontological politics. Through an examination of the complex relationship between ethics and politics in his work, this paper explores Levinas' notion of subjectivity as "otherwise than political," shaped by heteronomous responsibility and a critique of totality. By re-evaluating the ethico-political dimensions of his thought, the paper challenges the prevalent view of Levinas as apolitical, positioning him instead as a thinker whose understanding of subjectivity redefines the political as fundamentally ethical.

Keywords: Levinas, ethics, politics, subjectivity, totality, heteronomy, responsibility.

Simon Critchley (2004) observes that "Politics is an Achilles heel in Levinas" (p. 172). This observation echoes the renewed interest in the question of the 'political' in the otherwise celebrated ethical philosophy of Levinas. Scholars have made various attempts to bring to the limelight the 'political' in Levinas. However, Levinas' ambivalent treatment (of the question) of the political and the palpable ethical overtones of his thought resulted in a variegated reading of the political. Levinas places significant emphasis on the issue of politics, particularly while examining the challenging context of his era. He identifies the time he lived through as suffering from a political failure, the foundations of which he traces back to the shortcomings of the Western philosophical tradition. For him, these shortcomings are symptomatic of a philosophy that seeks to encompass everything within a totalizing framework. As a response, with his appeal to an 'ethical turn' in philosophy, Levinas calls for a shift away from the philosophy of totalization and its politics of ontology. However, this endeavour does not entirely dismiss the notion of the political. The resurgence of the political is visible in Levinas' problematizing the notion of justice and the 'third party' – *le tiers* (2013, p. 234, 1998, p. 16), which brings to light the ambiguity in decoding his idea of the political. This situation has led to diverse interpretations regarding the meaning and role of the political in Levinas' works.

This paper aims to revisit Levinas' notion of the 'political'. This approach might be considered a re-reading as much as it takes issue with standard readings of Levinas, which either consider him as an apolitical thinker or attempt to posit his political thought as an

*Department of Humanities and Social Sciences, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India; e-mail: antomsfs@gmail.com; ORCID ID: https://orcid.org/0009-0009-7003-3784

**Department of Humanities and Social Sciences, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India

ethical politics. Against the critics who contend that Levinas' philosophy is predominantly apolitical, this paper argues for a strong sense of the political in his philosophy by re-articulating its genesis and development as a response to his socio-political and cultural milieu.

In decoding the political in Levinas, the paper draws broadly on 'three waves' of scholarship that deal with the engagement of ethics and politics in his works, which ground our inquiry in the ongoing scholarly debate. The three popular strands of interpretation could be referred to as polemicists, who assert the antithetical nature of ethics and politics; apologists, who advocate for an ethical politics or a politics that is deduced from ethics; and deconstructionists, who advocate for a 'messianic politics', calling for a 'passage' from ethics to politics. By bringing the complexity of decoding Levinas' political in the context of the third party, the paper identifies the inadequacy of these readings. It takes issue with the primary assumption upon which these readings construe their thesis, which sees ethics and politics as two separate realms in contradiction with each other that they then attempt to resolve differently.

The paper advances an analysis of subjectivity in a threefold movement to decode the meaning of the political in Levinas. The first movement of subjectivity in Levinas presents a radical critique of the political as a movement away from totality, which suggests his deeper critique of the politics of totalitarianism. The first movement serves as the basis towards subjectivity as founded in heteronomous responsibility to the other which suggests the meaning of the political as a critique. This second movement of subjectivity can be considered as an 'otherwise than political' that moves away from the subjectivity predicated on rational, autonomous, and free subjects. These first two moments present the sense of the political in the negative register as an ethical critique. The third movement that introduces the complexity of the third party is analyzed in the paper as a positive 'return of the political.' The paper takes issue with the standard readings of ethico-political encounter, which consider the interaction of these movements encompassing the broader Levinasian project as a philosophical problem. This philosophical problem in Levinas' scholarship is rooted in a simplistic assumption that treats subjectivity's ethical and political moments as contradictions, disregarding the first two critical movements referred to as 'otherwise than ontological politics and otherwise than political.'

This paper argues that the encounter of ethics and politics could be better understood as rooted in an 'essential ambiguity' rather than a theoretical problem that requires philosophical overcoming. The Levinasian ideas of 'relation without relation' and 'the saying and the said' offer a hermeneutic key to rethink ethical-political interaction as paradoxically intertwined in its separation and relation. This reading, on the one hand, insightfully binds the imposing contradictions in Levinas' oeuvre in a manner that is sympathetic to his larger philosophical project and, on the other, offers a new way of thinking about ethics and politics that is neither reducible to one another nor contradictory binaries.

The paper is organized into three sections. The first section examines the possibility of the political in Levinas by posing his philosophical project as a response to his socio-political and cultural milieu. The second section posits the political in Levinas as an 'otherwise than' by explaining the three movements of subjectivity, taking a cue from his critique of Totality

(Politics of ontology), the ethical aspect of the heteronomous responsibility to the other, and the problem of the 'third party'. The third section brings these moments together as the problem of ethical-political interaction in Levinas to be understood in its 'essential ambiguity' by exploring the key notions of 'relation without relation' and 'saying and said,' showing their significance in making sense of Levinas' philosophical project.

1. Situating the Political in Levinas

Any attempt to understand the political in Levinas poses a host of problems. Levinas does not provide any systematic account of the political, which should not lead us to superficial conclusions and label him as an apolitical thinker. If so, this would be an injustice to Levinas' thinking because he does not give a systematic ethical theory in the traditional sense either. His accounts clearly show that he is not concerned with the theorization of ethics but rather with finding 'the meaning of ethics' (Levinas, 1985, p. 95). Taking a clue from Levinas' critical engagement with ethics as a non-allergic relationship to the other justifies seeking the meaning of the political in Levinas on a deeper level, not in the sense of a political theory rather as an otherwise than political. Thus, understanding the meaning of the political poses three crucial questions: What meaning does the 'political' assume in Levinas' philosophical project? How should we unfold the import of his ideas in the socio-political-cultural milieu they inhabit? What is the upshot of this inquiry? Responses to these questions help us understand the nature of the relationship between ethics and politics and how, in turn, this shapes the idea of ethico-political subjectivity in Levinas' work.

In a recent article titled "Levinas and the Question of Politics," Robert Froese (2020) emphasizes the necessity of raising the political question within Levinas' broader engagement with philosophy (p. 1). This underscores the need for a hermeneutic task to unravel the political aspect of Levinas' thought and grasp the potential emergence of ethico-political subjectivity in his framework. The classification of Levinas' scholarship into three waves aids this exploration (Atterton & Calarco, 2010, p. x). The first wave tended to overlook the political dimension, centring instead on 'the face' in *Totality and Infinity*, thereby characterizing Levinas predominantly as an ethicist (Bernasconi, 1999, p. 76). The second wave, prompted by Derrida's critique, investigates the political implications of Levinas' oeuvre, particularly the interrelation between ethics and politics, in the context of the third party (Atterton & Calarco, 2010, p. x). The Third Wave that emerged in the 1990s, exemplified by scholars like Bernasconi and Critchley, focuses on the ethical-political interface, which presents three primary responses to the political question.

The first response, upheld by sympathetic interpreters, accentuates the ethical domain's primacy, resisting any shift towards politics, citing Levinas' dictum that "politics left to itself is Tyrannical" (Levinas, 2013, p. 300). Conversely, the second perspective advocating for a deductive socio-political framework rooted in Levinasian ethics calls for a passage from ethics to politics through the introduction of the 'third party' (Gibbs, 1997, p. 57). Deconstructionist scholars, influenced by Derrida, discern a crucial gap between ethics and politics in Levinas, emphasizing the need to bridge this gap to facilitate a passage from ethics to politics, as articulated by Critchley (1992), Campbell (1998), and Hagglund (2004). Besides these, in recent times, Marxist scholars have engaged in a critical dialogue with Levinas' ideas, focusing on his purported avoidance of "the basic philosophical fact that

identity can never be posited directly and must appear within a field of differentiation" (Froese, 2020, p. 5). So, the relationship between ethics and politics in Levinas has been a subject of frustration, puzzlement, and, to a significant extent, a catalyst for scholarly reticence and academic indifference.

The paper takes issue with these readings of the political in Levinas, which ground their arguments on the foundational distinction between ethics and politics, viewing them as discrete domains necessitating mediation for a transition between them. However, departing from this viewpoint, the paper draws upon Bernasconi (1999) for inspiration. It aligns more closely with Levinas' broader philosophical stance, which regards ethics and politics as intricately interconnected yet distinguishable. Following Bernasconi's lead, Madeleine Fagan posits a compelling case for a poststructuralist reinterpretation of Critchley, Hagglund, and Campbell, critiquing their advocacy for a mandated separation of ethics and politics and instead advocating for inseparability of the two in Levinas (2009, pp. 5-22). As a first step towards decoding the meaning of the political in Levinas, we suggest Levinas' project as a philosophical response to a distinctive socio-political-cultural milieu.

A deeper reading of Levinas would bring us closer to the understanding that his philosophy predominantly responds to a socio-political situation, with its debilitating experience of the terror of the Holocaust and the world wars. From the very beginning of his childhood, Levinas had to endure the political tyranny of the Bolsheviks. As a student philosopher, he had to personally go through the trauma of the draconian visage of Auschwitz and the Second World War. This existential predicament of fear, misery, violence, and persecution as a Jewish refugee sets the tone for his philosophical inquiry, which is evident from his major works. In the preface to *Totality and Infinity*, he remarks: "Everyone will readily agree that it is of highest importance to know whether we are not duped by morality" (Levinas, 2013, p. 21). In this work, we notice a sense of urgency to prevent war propagated and founded on a philosophy of totalization.

His meditations about war in a more amplified way can be found in *Otherwise than Being or Beyond Essence*, where he affirms: "War is the deed or the drama of the essence's interest" (Levinas, 1998, p. 4). Reflecting on Levinas' insights on war and its relationship to the other, Smith (2005) observes that for Levinas, "war leaves nothing exterior to it and changes the other to the Same. Is this not another way of expressing the Hitlerian directive, the Gleichschaltung, a command the entire Reich was to think and act in unison?" (p. 113) However, it should be noted that as early as 1934, in his rather unpopular essay reflecting on the doctrine of Hitlerism, Levinas (1990) makes his robust analysis that "... the source of bloody barbarianism of National Socialism lies not in some contingent anomaly within human reasoning or in some accidental ideological misunderstanding." He instead traces its origin in the "essential possibility of elemental evil into which we can be led by logic and against which western philosophy has not sufficiently insured itself" (Levinas, 1990, p. 63). In several of his works, Levinas acknowledges the inextricable relationship between war, the Holocaust and the philosophy of Totalization characteristic of the Western philosophical tradition and its effects on shaping his philosophical enquiry (Levinas, 1996; 1998). He observes that the feeling and the memory of the Nazi horror have influenced his entire life (Levinas, 1990). The Holocaust, for Levinas, was nothing but an outgrowth of the philosophy of totalization,

the impact of which makes an indelible mark in his life and thought. As Michael Smith (2005) remarks:

> The rise of anti-semitism in the 1930s, the threat of totalitarianism, and the Holocaust itself not only 'mark' Levinas's work, but they also set in motion a complex intertwining of his philosophical reflection with the state of the world in which it unfolded (p. 109).

The above discussion suggests a solid basis for our argument that Levinas' philosophy is indeed an answer to the debilitating socio-political-cultural milieu. Therefore, to comprehend the political in Levinas, we may have to follow the contours of the experiences and encounters he endured during his early years. In this sense, it involves re-reading Levinas' basic project as a response to his socio-political-cultural engenderment. He considers this very socio-political crisis of his time as the existential predicament that shapes his otherwise than Western philosophical thinking, particularly in his critique of ontology and totality. For him, the root of such socio-political violence is deeply embedded in Western philosophy's general thought process. As Caygill (2002) observes, Levinas "... unambiguously aligned his philosophy with the thinking of political, or, more precisely, with the thinking of political horror" (p. 5).

2. Re-reading of Subjectivity (ethico-political) as 'Otherwise than Political'

The discussion above suggests the importance of situating our inquiry in the context of his philosophical project as a response to the political horror of totalitarianism, violence and war. In order to understand the meaning of the political in Levinas we offer a re-reading of subjectivity as otherwise than political in its threefold movement, namely, subjectivity as a movement away from totality; subjectivity as founded in heteronomous responsibility; and subjectivity as ethical-political encounter.

A. Subjectivity as a Movement Away from Totality

Levinas is known for his severe critique of the Western philosophical tradition, a necessary edifice of his philosophical project. However, this staunch critique gained scant attention until Derrida (1974) discussed it in his "Violence and Metaphysics". According to Levinas, enquiring into the 'why' of the debilitating visage of the post-war era results in identifying the totalizing inclinations of Western thought as foundational for political tyranny. In *Totality and Infinity*, Levinas raises a fundamental question; "are we not duped by morality?" (2013, p. 21). This could be inextricably connected to another fundamental question: Are we not duped by politics? The focal point of our philosophical deliberations would be on the 'how' of unfolding of subjectivity via Levinas' critique of the Western philosophical tradition rooted in the hegemony of ontology and philosophy of totality.

In *Totality and Infinity*, Levinas delineates and inaugurates a new form of interrogation and rethinking, constituting the very edifice of Western philosophical tradition. He observes that Western philosophy has predominantly focused on ontology, which entails the quest to

understand the fundamental nature of what exists. This orientation, he argues, tends to simplify the diversity of existence by reducing it to a uniform and homogenized concept, a process he succinctly captures as "a reduction of the Other to the Same" (Levinas, 2013, p. 43). Within this philosophical tradition exists a corresponding epistemological perspective where objects of cognition are interpreted as objects inherent to the realm of consciousness. Essentially, this tradition can be described as "suppressing or reducing all forms of otherness by transmuting them into the same" (Critchley, 2002, p. 16). Levinas' staunch critique stresses the pivotal position totality assumes in shaping the contours of politics and warfare and the broader fabric of societal institutions within the Western philosophical tradition, which embodies a systematic manifestation of 'Universal Reason', seeking to encompass the full spectrum of human experience within a singular, unifying theory or practice. In this light, Levinas presents totality as emblematic of what he critically refers to as the "tyranny of the Same" (Kleinberg, 1999, p. 45). This inclination arises from the inherent tendency of the self to assimilate all aspects of existence into its framework, thereby diminishing the distinctiveness of the other. Levinas contends that this inclination is a characteristic shared by the dominant tradition in Western philosophy, epitomized by thinkers like Heidegger, wherein individual and concrete beings are subsumed beneath the abstract concept of impersonal Being (Kleinberg, 1999, pp. 47–48). In this critical assessment, Levinas identifies Heidegger as the key figure who championed the logic of the philosophy of totality in its breadth and width. He acknowledges Heidegger's contribution to philosophy, especially his critique of the Cartesian 'thinking subject' and the subject-object dichotomy. Nevertheless, he affirms that, despite his efforts in decentring the subject with his novel historicized phenomenology of Being, Heidegger was not entirely free from the sway of the philosophy of totality. Levinas' critique focuses on the primacy that Heidegger endorses in the Ontology of Being, which provides meaning and structure to everything that is. For Levinas, Heidegger's critique of the primacy of the subject is valid. However, he argues that it does not sufficiently address the question of alterity, which is overshadowed by the Ontology of Being in its all-encompassing structure of totality that reduces everything unto itself (2013, pp. 45–46).

In his turn, he argues for the need to move away from the general climate of the Western philosophical tradition as a philosophy of totality. He seeks an alternative rethinking of Western philosophy, oriented by a radical turn to the other. Levinas questions the general obsession with totality and universalizing essence, which is foundational to all forms of thematization and representation. Instead, he asserts the need to pay attention to the concepts of infinity and transcendence. However, Levinas does not entirely do away with totality. While being a staunch critic of totality, he acknowledges its importance as necessary for society to exist and function, which becomes apparent in our discussions on the third party.

What is evident in Levinas' argument is a double movement: one, a movement away from totality with its potential to give rise to devastating and disastrous consequences, and two, a movement towards the other founded on heteronomous responsibility to the other. A double movement of 'away and towards' is the movement away from the political (politics of ontology), which is necessary for a movement towards ethical (heteronomous responsibility to the other). He observes that his critique of Totality "came in effect, after a political

experience (Holocaust) that we have not forgotten" (Levinas, 1985, p. 73). Levinas sees the foundations of political tyranny and violence in such a philosophy of totality. This would force Levinas to leave the climate of this philosophical totality with its political tyranny, which serves as the basis for our argument for decoding the political in Levinas as otherwise than political. From Levinas' critique of totality, we can tease out the meaning of the political in the negative register, a movement away from totalitarian politics.

Gavin Rae (2016), in his readings of Levinas' critique of totality, also suggests that Levinas' staunch critique of Western philosophical tradition is more nuanced than it appears to be at the very outset. For Rae, the nuances consist in how Levinas enshrouds the political in his critique of totality, which calls for a movement away from totalitarian politics. Rae argues that Levinas identifies "the problem of politics" as a primordial problem present throughout the history of Western philosophy (Rae, 2016, p. 141). The discussions suggest the meaning of the political masquerading under the critique of totality as otherwise than political in its movement away from ontological politics as a necessary condition for a movement towards ethical responsibility in the face-to-face. In the following discussion, let us look at the unfolding of the political in exploring the movement of subjectivity as founded on heteronomous responsibility.

B. Subjectivity as Founded in Heteronomous Responsibility

To understand further how the movement of subjectivity as founded in heteronomous responsibility unfolds the understanding of the political in Levinas, we will briefly look into the idea of heteronomous responsibility, situating it within the problematic of subjectivity as conceived by modern political or liberal political thought where ideas like autonomy, freedom, individualism and equality are fundamental. Tahmasebi-Birgani notes that Levinas' text hardly engages with politics as a distinct philosophical tradition; instead, he fundamentally reframes the issue so that politics can no longer be assumed to be excluded from the asymmetry of one's responsibility for justice demanded in the face of the other. His writings also raise the topic of politics when discussing other subjects, including his views on justice and law, the subject, freedom, responsibility, peace and war, and the rationale for the existence of the state (Tahmasebi-Birgani, 2014, p. 20).

In modern political philosophy, the notions of autonomy and freedom occupy pivotal positions. Opposing this, Levinas seeks new foundations for subjectivity otherwise than the modern political framework, turning towards the other or the phenomenology of the face. In Levinas' view, Hobbes undervalues the ethical relationship between self and the other, essential for establishing society and peace. Levinas argues that Hobbes' understanding of the state of nature and, subsequently, his political framework is based on the central undertaking of the Western metaphysical tradition, the foundation of which is the idea of an autonomous, rational, free ego solely interested in its self-preservation. Borrowing Spinoza's vocabulary, Levinas describes the ego as "conatus essendi," which means the "struggle of living" or "effort of being" (1998, p. 127). This way of thinking inadvertently creates the conditions for an attitude of dominance, violence, tyranny, and warfare. Levinas questions the Hobbesian state of nature as a "war of all against all" (2013, pp. 21–30). Another essential aspect of Levinas' critique of subjectivity is constituted by the phrase "freedom

called into question by the other" or "freedom put into question by the other" (Levinas, 2013, p. 85), which takes our discussion further to the notion of heteronomous. To understand further how the movement of subjectivity as founded in heteronomous responsibility unfolds the understanding of the political in Levinas, we will briefly look into the idea of heteronomous responsibility, situating it within the problematic of subjectivity as conceived by modern political or liberal political thought where ideas like autonomy, freedom, individualism and equality are fundamental. Tahmasebi-Birgani notes that Levinas' text hardly engages with politics as a distinct philosophical tradition; instead, he fundamentally reframes the issue so that politics can no longer be assumed to be excluded from the asymmetry of one's responsibility for justice demanded in the face of the other. His writings also raise the topic of politics when discussing other subjects, including his views on justice and law, the subject, freedom, responsibility, peace and war, and the rationale for the existence of the state (Tahmasebi-Birgani, 2014, p. 20).

In modern political philosophy, the notions of autonomy and freedom occupy pivotal positions. Opposing this, Levinas seeks new foundations for subjectivity otherwise than the modern political framework, turning towards the other or the phenomenology of the face. In Levinas' view, Hobbes undervalues the ethical relationship between self and the other, essential for establishing society and peace. Levinas argues that Hobbes' understanding of the state of nature and, subsequently, his political framework is based on the central undertaking of the Western metaphysical tradition, the foundation of which is the idea of an autonomous, rational, free ego solely interested in its self-preservation. Borrowing Spinoza's vocabulary, Levinas describes the ego as "conatus essendi," which means the "struggle of living" or "effort of being" (1998, p. 127). This way of thinking inadvertently creates the conditions for an attitude of dominance, violence, tyranny, and warfare. Levinas questions the Hobbesian state of nature as a "war of all against all" (2013, pp. 21-30). Another essential aspect of Levinas' critique of subjectivity is constituted by the phrase "freedom called into question by the other" or "freedom put into question by the other" (Levinas, 2013, p. 85), which takes our discussion further to the notion of heteronomous[1] responsibility as manifested in the face of the other. For Levinas, this 'freedom called into question by the other' opposes the philosophical tradition that views freedom as autonomy. Levinas argues that freedom is fundamentally rooted in heteronomy. He contends that without the ethical encounter with the face of the other, the self would lack the impetus to question itself and its actions. This view challenges the traditional understanding of the modern subject, which, from Descartes to Kant and beyond, often conceives of freedom as rational self-governance. He does not intend to negate the significance of freedom but seeks to provide a radically different explanation and justification for it. This becomes clearer when discussing the ethical investiture of responsibility as preceding freedom.

Modernity has accustomed us to view responsibility as being accountable for one's actions and statements in front of others, which is predicated upon the autonomous, rational, free subject. Burggraeve rightly observes that "since the Enlightenment, anyone viewing the world through a Western lens has become used to thinking of responsibility as the

[1]Here, the usage of 'heteronomous responsibility' is borrowed from Roger Burggraeve's interpretation of Levinas to emphasize the passivity that is characteristic of this unique understanding of responsibility. For more detail, please see (Burggraeve 2009).

extension of, and even the synonym for, autonomy and freedom" (2009, p. 66). Levinas, however, offers a significantly more radical perspective on responsibility that goes beyond the notion of active commitment to oneself. John Caputo shares this viewpoint with Levinas, particularly when he characterizes responsibility as "to give up your Greek autonomy and agency and to submit to the 'jewgreek' heteronomy and patience" (1993, p. 125). Levinas' philosophical insight constitutes a new conception of responsibility that may bear a renewed understanding of [political] subjectivity as an otherwise than. As he claims, "The unlimited responsibility in which I find myself comes from the hither side of my freedom, from a "prior to every memory," an "ulterior to every accomplishment," from the non-present par excellence, the non-original, the anarchical, prior to or beyond essence" (1998, p. 10). As Dimitrova observes, "My responsibility is my being chosen, my 'birthright' – I am the first who can respond to the call of the Other in a given situation and I am thus unique and irreplaceable in it" (2016, p. 137). This underscores the uniqueness of my responsibility that is impossible to transfer to something neuter or some institutions.

Levinas posits that the profound alterity of the other and the ensuing demand it places upon me render the "I" responsible, both to and for the other. This responsibility is absolute and unconditional, yet it does not originate from within "I" but precedes me. To clarify, it precedes me in the sense that it has its origins in someone who precedes my existence, existing in a pre-original state. This responsibility harkens back to an irretrievable past, lying in a pre-ontological realm (Levinas, 1998, p. 78). It is rooted in a time that predates my freedom and even precedes the commencement of my existence. This responsibility emerges from an absolute past that lacks any specific subject (Levinas, 2013, pp. 56, 85, 89); this means that, as long as one exists, one is inherently bound to be responsible for the other. Thus, according to Levinas, this responsibility takes precedence over freedom. Levinas emphasizes that one does not voluntarily choose to be responsible; instead, responsibility is imposed by the other in a unique ethical encounter where one discovers one's own uniqueness. This notion aligns with Levinas' citation of Dostoyevsky's *Brothers Karamazov*, asserting, "We are all responsible for everyone else, but I am more responsible than all the others" (1985, 98). This critique provides a strong case for Levinas' conception of subjectivity in opposition to the liberal political framework as an otherwise-than-political constituted by heteronomous responsibility for the other. The proposed radical shift in philosophizing from the primacy accorded to the self-interested autonomous, free subject towards a philosophy predicated on heteronomous responsibility to the other provides breakthrough insights towards a radical rethinking of the political, which suggests the meaning of the political as something to be overcome in the negative register.

C. The Ethico-Political Encounter

Taking a cue from previously discussed discourse on subjectivity that appears in Levinas as otherwise than totalizing political towards a radical turn to the other as an alternative to the otherwise rationalistic mask of Western philosophy, we are left with the following understanding of the political in Levinas. First of all, the decoding of the political in Levinas implies a movement away from totality, which implies an escape from the totalizing political. Secondly, this movement away from calls for a movement towards heteronomous responsibility, which implies a radical critique of the political status quo of the modern era

predicated on a rational, autonomous, free-thinking ego. This analysis so far leaves us with an understanding of the political in Levinas in the negative register, something that he is deeply suspicious of and wants to overcome. This subjectivity, founded on asymmetrical heteronomous responsibility and applied on a one-to-one basis without reference to reciprocity, poses many questions and raises some important concerns. One of the main worries is that the idea of heteronomous responsibility based on an asymmetrical dyadic (one-to-one) link between the "I"/Self and the "other" fails to treat everyone fairly. This is a threat to the administration of justice[2] (accountability/responsibility) in our everyday life. Levinas does not limit his ethics to the intimate one-to-one relation between the self and the other. He argues, "The face that looks at it places itself in the full light of the public order, even if I draw back from it to seek with the interlocutor the complicity of a private relation and a clandestiny" (Levinas, 2013, p. 213). Levinas proposes the concept of the third party (*le tiers*) to allude to the existence of 'other' other, in order to address the socio-political question, or the issue of justice (2013, p. 234, 1998, p. 16).

The notion of the third party plays a key role in facilitating a return of the political in Levinas. However, this return of the political in Levinas does not mean a return of the politics of ontology or the rationalistic subject that he calls into question in his critique of totality. Peperzak (1993) observes that the presence of a third party leads to a division of my ethical responsibility (p. 229). In Levinas' framework, the third party serves a dual purpose. The third party is important for Levinas as it serves as a reminder to the Subject (the one responsible) that their ethical responsibility toward the other (another individual) should not be limited to just that specific other but should extend to encompass other others as well (the third). In other words, it emphasizes the need for an inclusive sense of responsibility. Secondly, introducing a third party allows the Subject to recover from the seemingly boundless and overwhelming nature of their infinite responsibility toward the other. When the third party becomes part of the ethical equation, the Subject's responsibility becomes distributed or shared somehow, alleviating the unrelenting burden of infinite responsibility that might otherwise be experienced. This dual function of the third party in Levinas' philosophy highlights the complex nature of ethical responsibility, which poses serious questions about the role of ethics and politics in Levinas. Does this openness to the 'third' in any way limit/negate my responsibility to the other? Does the beginning of politics and justice do away with ethics in Levinas? Is ethics compatible with politics in Levinas?

Levinas defines justice as "the limit of responsibility and birth of the question" (1998, p. 157). On the other hand, he asserts that "in no way is justice a degradation of obsession, a degeneration of the for-the-other, a diminution, a limitation of anarchic responsibility" (p. 159). This presents the problem of conceiving a relationship between ethics and politics without limiting the anarchic responsibility. The problem is exacerbated in the wake of the contradictory accounts Levinas provides regarding the arrival of the third party. This has led to different readings of the third party, which has diverse consequences for understanding the relationship between ethics and politics in Levinas. In what follows, we take issue with

[2] "Accountability" could correctly be used as a synonym of responsibility in the context of justice with the appearance of the third party. "Accountability" may be suitable as far as "the I" has to compare different cases and make "accounts" concerning them.

these readings to explore the possibility of decoding the meaning of the political in Levinas as otherwise than political.

3. Paradoxical intertwining of ethics and politics

The plethora of questions that stem from the discussions on the notion of the 'Third' has succumbed to broader disputes given different phenomenological accounts of the third party in Levinas. Based on their readings of the third, in answering the question, when does the third party exist? Scholars have developed respective theories to understand the relationship between ethics and politics. One of the responses can be drawn from the popular readings of *Totality and Infinity*, which privileges the face-to-face relation. This reading claims that: "the Third's arrival is not spontaneous with or contained within, the Self's face to face encounter with the other, but it is a distinct and separate event that contextualizes it" (Fagan, 2009, p. 31). According to this reading, the political must be deduced from the logically prior original ethical relation. This is based on the linear phenomenological progression in *Totality and Infinity* from interiority to exteriority to the emergence of a socio-political world (Levinas, 2013, p. 300). This reading of the third is rooted in an assumption that considers ethics and politics (Justice) as two realms of the relation to the other, which suggests the possibility of deducing or overcoming the gap between them (Fagan, 2009, p. 15). This is the basis of Žižek's reading of Levinas, which disregards the potential for political theorization within Levinas' framework. He observes that "[f]ar from preaching an easy grounding of politics in the ethics of the respect and responsibility for the Other, Levinas rather insists on their absolute incompatibility, on the gap separating the two dimensions" (Žižek, 2006, p. 149).

The second reading about the third party obviates the need for a deduction of the political from the ethical since the ethical is already political. This reading is based on Levinas' argument that "the third party looks at me in the eyes of the other–language is justice" (2013, p. 213). While the social deductive reading commits the violence of reducing politics to the ethical, the second reading in its attempt to avoid the inadequacy of the first theory by advocating the co-existence of ethics and politics, dilutes the primacy of separation between ethics and politics, which guarantees their uniqueness. In contrast to these readings, we argue for a re-reading of the third as a paradoxical intertwining that facilitates a separation and relation, escaping the assimilative and reductive logic that Levinas' philosophy seeks to overcome.

We could derive the meaning of the paradoxical intertwining by looking at Levinas' notion of 'relation without relation'. Levinas introduces this notion in problematizing the relationship between self and other. The 'relation without relation' has three essential characteristics. First, the relation is characterized by an essential separation that distinguishes the self and the other. However, the separation does not indicate an opposition between the two. Second, for Levinas, 'the without relation' of the absolute does not place the other in an isolated realm; "these bonds do not unite the same and the other into a whole" (2013, p. 48). In this specific context, the concept of absoluteness serves to transcend the limitations inherent in relationality. It avoids attributing an unattainable, distinct quality to the other party. Instead, the term 'absolute' challenges the idea of defining a relationship explicitly and that of defining it through the relationship itself. The notion of the absolute other does

not imply an otherness removed or isolated from existence. Instead, it signifies an otherness that exists beyond the confines of categorization. Its absoluteness indicates resistance to being confined or determined by pre-established categories (2013, p. 71). In other words, its absoluteness means resistance to determination. Two characteristic features of this relation are the key moments of separation and relation. There is an essential separation that is maintained between the self and the other, which guarantees the self-sufficiency of both. The self is simultaneously called into an infinite responsibility for the other without reducing it to the same in its relationship with the other (Levinas, 2013). A closer look at the notion of 'relation without a relation' in Levinas shows how such a reading concords with his basic philosophical task, which requires: "a description and defence of subjectivity, but also an account of alterity which does not reduce the Other to the Same; and finally, it needs some means of accounting for the relation between Same and Other that does not effectively abolish either" (Davis, 1996, p. 41).

Reappropriating this task in the context of the relation between ethics and politics could be read as protecting ethics from the aggressions of politics; to look for the possibilities of its appearance in politics and, finally, to provide some means of accounting for the relation between ethics and politics that does not effectively abolish either. Levinas' philosophical task and the analysis of the notion of 'relation without a relation' thus establishes a paradoxical intertwining of ethics and politics as consistent and faithful to his philosophical task. Levinas' embracing of paradoxes and essential ambiguity can also be seen as one way to go beyond traditional philosophical thinking based on the dialectics of binaries, which conceives the relationship between ethics and politics in terms of essential opposition and separation. Such an alternative reading of Levinas provides an insightful understanding of the relationship between ethics and politics as 'comparing the incomparable', sustaining the paradoxical intertwinement between the two without dissolving into each other or to a third whole. Levinas often presents this paradoxical intertwinement between ethics and politics (justice) as an "essential ambiguity" (2001, p. 230).

The traces of this essential ambiguity can be found in Levinas' description of the saying and the said. Simon Critchley points out that one of the significant achievements of *Otherwise than Being or beyond Essence* is the structure of 'the saying and the said' that is how the ethical could signify within the ontological language (1992, p. 164). The objective here is not to give an exhaustive account of the saying and the said in Levinas but to briefly see how this relationship leaves the traces of the paradoxical intertwining supporting our analysis. The saying for, Levinas refers to the ethical excess, 'the overflow of conjunction' or 'non-thematizable', which escapes every form of thematization (1998, p. 11). The said refers to that aspect of language that is 'fixed, assembled in tale, is synchronized, presented, lends itself to a noun, receives a title' (1998, p. 42). Levinas poses this relationship in the mode of a betrayal – the betrayal of the saying in the said. For Levinas, this betrayal of the saying in the said is necessary for as much as Levinas' whole philosophical project unfolds around the one question: how can my exposure to the Other be said or described in philosophical exposition without finally betraying this saying? (Critchley, 2002, p. 18).

For Levinas, the saying needs to be thematized in the said in order to manifest itself, which requires the return of the ontology. However, this return of ontology in the betrayal

of the saying in the said does not take away the non-totalisable or non-thematizable excess of the saying. Levinas expresses this paradox in the phrase 'a comparison of incomparables' in *Otherwise than Being or beyond Essence*, which expresses key insight that ontology needs to return because of the political but in such a way that it remains challenged by ethical asymmetry (1998, p. 158). This key insight lends itself in the same way that saying needs to be Said and therefore betrayed, but always returns behind the said, leaving its traces as an excess. Similarly, the relation with the other needs to be understood, categorized, and thematized in the political to avoid it being fixated on interpersonal relationships, which brings in the question of justice/politics – which, at the same time, does not compromise the primacy of the anarchic responsibility. Accordingly, we could trace a paradoxical intertwining relationship of the saying and the said in its movement made possible by the betrayal of the saying in the said beyond logos or essence, "to show the signification proper to the saying on the hither side of the thematization of the said" (Levinas, 1998, p. 43). Nonetheless, what we should take into account is that even though Levinas maintains that "philosophy remains the servant of the Saying", there is a need to synchronize the Saying into the Said because "philosophy is called to thought by justice" (1998, p. 162). For Levinas, "there is a question of the said and being, only because saying or responsibility requires justice" (1998, p.45). This is because a responsibility that would exclude all others in the exclusivity of the one-for-the-other would necessarily lose sight of the meaning of the for-others. The relation between the unthematized Saying and the thematized Said, in its essential ambiguity characterized by its betrayal of the Saying in the Said reflects the paradoxical intertwinement between ethics and politics/justice as distinct yet intertwining ways of the exposure to the other.

Thus, while ethics refers to the one dimension of this relationship to the other characterized by infinite openness, asymmetry, and resistance to closure, it finds its essence in the "saying" – a pre-conceptual, unthematized response to the Other's call that grounds meaning and embodies an ongoing relational movement. In contrast, politics, embodied in the "said," involves structure, determinacy, and thematization, manifesting the other dimension of the relationship to the other. While the saying precedes and grounds the said, it also depends on the said for its manifestation. This interplay ensures that the ethical dimension of the saying continually critiques and exceeds the finite structures of the said, preventing ethics from being fully absorbed into politics while simultaneously holding politics accountable to the ethical demand.

This paradoxical dynamic reveals the simultaneous separation and relation between ethics and politics. Ethics, with its infinite and asymmetric demand, resists being reduced to the calculative, symmetrical systems of politics. However, politics is necessary to make the ethical practical and intelligible through institutions that promote justice. Ethics remains an enduring critique of political totality, ensuring that political structures do not become complacent or detached from their ethical grounding. At the same time, politics preserves the relevance of ethics, preventing it from becoming an abstract dyadic relationship by unfolding the ethics/saying within the Said/Political. This dual relationship resists any resolution into a simple unity or dichotomy, highlighting that politics must remain perpetually accountable to the ethical. Thus, Levinas' framework underscores the paradoxical intertwining of the ethical and the political in their separation and relation, with ethics requiring justice/politics

for its manifestation while politics is constantly interrupted by the ethical in its exposure to the other.

Our attempt to understand the nature of the political in Levinas, within the broader debate on the relationship between ethics and politics in his scholarship, leads us to the following conclusions. This analysis suggests the need to move beyond simplistic readings of Levinas that often reduce him to an apolitical thinker with little to contribute to political theorizing. In contrast, this paper decodes the meaning of the political as "otherwise than political," highlighting a paradoxical intertwining of the ethical and the political in Levinas' thought.

Situating this discourse as an unfolding of subjectivity away from totality and toward heteronomous responsibility and ethico-political encounters in the presence of the third party underscores the importance of deeper engagement with the political in Levinas. The critical exploration of subjectivity as a movement away from totality reveals the complexity of Levinas' critique of totalitarianism, rooted in a politics of ontology. Simultaneously, the notion of subjectivity as a movement toward heteronomous responsibility provides a radical critique of modern political theories based on the autonomous, rational, and free individual. Together these critiques unfold the meaning of the political in its negative shade as something to be overcome. The arrival of the third party further illustrates the essential ambiguities involved in understanding the relationship between ethics and politics in Levinas' scholarship. It also marks the "return of the political" in Levinas, though not in the totalizing sense he critiques and seeks to move away. Reflections on the third party as an unfolding of the paradoxical intertwining of ethics and politics, highlight the dynamic interplay of separation and relation. This perspective offers a nuanced understanding of the political in Levinas as "otherwise than political", which moves beyond the two competing approaches to ethics and politics in Levinas: one that views them as contradictory and oppositional, reducing politics to ethics, and another that advocates for a necessary and linear mediation from ethics to politics. Instead, it provides an alternative understanding of ethics and politics as paradoxically intertwined, preserving their separation and relation, which maintains the primacy of the ethical without reducing or subsuming politics. Hence, although Levinas does not provide a systematic political theory, he provides necessary insights for political theorization as interrogations calling to question our responses to radicalize political subjectivity as an otherwise than. As Critchley aptly states, "Ethics without politics is empty, and politics without ethics is blind" (2007, p. 13).

4. Concluding Remarks

In examining Levinas' thought through the lens of ethico-political encounters, this paper has drawn out a new perspective that offers a decoding of the political in Levinas as otherwise than political (that exists distinct from conventional political paradigms). This reevaluation prompts a reconsideration and readdressing of the prevalent critique that deems Levinas' thought as inherently apolitical. His rigorous critique of the self-contained political domain raises a polemic against a politics divorced from an ethical commitment to the radical alterity of the other. Nevertheless, this paper seeks to secure Levinas' status as a socio-political thinker who argues for a subjectivity as 'otherwise than political' that traces the meaning of subjectivity as ethico-political rather than traditionally based on a totalizing

ontology that is not open to ethics. Indeed, his philosophical stance is fundamentally responsive to the socio-political milieu within which it is situated. Moreover, Levinas' critique of the Western philosophical tradition unfolds a notion of political subjectivity that departs from the totalizing inclinations inherent to this tradition, particularly its ontological underpinnings. This departure signifies a shift towards embracing heteronomous responsibility to the other as the driving force. Through this transformation, Levinas forges an ethical approach that resonates with political implications, upending conventional binaries. Finally, in light of the discussions and arguments, the paper opens up a new orientation to reconceive and decode the political in contemporary philosophical debates, as illuminated by ethico-political encounters (which call for an interwoven relationship between ethics and politics). This junction marks the beginning of a trajectory that beckons further exploration and research. As such, there lies a promising prospect for probing into the intricate connection between Levinas' philosophy and the political domain, potentially leading to an evolved ethico-political engagement attuned to the demands of the contemporary socio-political landscape.

References

Atterton, P., & Calarco, M. (2010). Editors' Introduction: The Third Wave of Levinas Scholarship. In P. Atterton, & M. Calarco (Eds.), *Radicalizing Levinas* (pp. ix–xviii). Albany: Suny Press.

Bernasconi, R. (1999, January). The third party: Levinas on the intersection of the ethical and the political. *Journal of the British Society for Phenomenology*, 30, 76–87.

Burggraeve, R. (2009). *Each other's Keeper?: Essays on ethics and the biblical wisdom of love*. Thrissur: Marymatha Publications.

Caputo, J. (1993). *Against ethics: Contributions to a poetics of obligation with constant reference to deconstruction*. Indianapolis: University Press.

Caygill, H. (2002). *Levinas and the political*. New York: Routledge.

Champbell, D. (1998). *National deconstruction: Violence, identity and justice in Bosnia*. Minneapolis: University of Minnesota Press.

Critchley, S. (1992). *The Ethics of deconstruction: Derrida and Levinas*. Edinburg: Edinburgh University Press.

Critchley, S. (2002). Introduction. In S. Critchley & R. Bernasconi (Eds.), *The Cambridge companion to Levinas* (pp. 1–32). Cambridge: Cambridge University Press.

Critchley, S. (2004, April). Five problems in Levinas's view of politics and the sketch of a solution to them. *Political Theory*, 32(2), 172–185.

Critchley, S. (2007). *Infinitely demanding: Ethics of commitment, politics of resistance*. London: Verso.

Davis, C. (1996). *Levinas: An introduction*. Cambridge: Polity Press.

Dimitrova, M. (2016) *Sociality and Justice – Toward Social Phenomenology*, Ibidem, Kindle.

Fagan, M. (2009). The inseparability of ethics and politics: Rethinking the third in Emmanuel Levinas. *Contemporary Political Theory*, 8, 5–22.

Froese, R. (2020). Levinas and the question of politics. *Contemporary Political Theory*, 19, 1–19.

Gibbs, R. (1997). Asymmetry and mutuality: Habermas and Levinas. *Philosophy and Social Criticism*, 23, 51–63.

Hagglund, M. (2004). The Necessity of discrimination: Disjoining Derrida and Levinas. *Diacritics*, 34(1), 40–71.

Kleinberg, E. (1999). Ethics beyond the body: Descartes and Heidegger in Emmanuel Levinas's *Totality and Infinity*. *Paroles gelées*, 17(2), 43–55.

Levinas, E. (1985). *Ethics and infinity: Conversations with Philippe Nemo*. (R. Cohen, Trans.) Pittsburgh: Duquesne University Press.

Levinas, E. (1990). Reflections on the philosophy of Hitlerism. *Critical Inquiry*, 64–71.

Levinas, E. (1990). Signature. In *Difficult Freedom: Essays on Judaism* (S. Hand, Trans.), pp. 291–95. Baltimore: The Johns Hopkins University Press.

Levinas, E. (1996). Nameless. In *Proper names* (M. B. Smith Trans.), pp. 119–125. Stanford University Press.

Levinas, E. (1998). *Otherwise Than Being or Beyond Essence*. (A. Lingis, Trans.) Pittsburgh: Duquesne University Press.

Levinas, E. (2001). Responsibility and substitution. In J. Robbins (Ed.), *Is it righteous to be* (G. Maureen, Trans.), pp. 227–233. Stanford: Stanford University Press.

Levinas, E. (2013). *Totality and infinity: An essay on exteriority*. (A. Lingis, Trans.) Pittsburgh: Duquesne University Press.

Peperzak, A. (1993). *To the other: An introduction to the philosophy of Emmanuel Levinas*. Indiana: Purdue Research Foundation.

Rae, G. (2016). *The problem of political foundations in Carl Schmitt and Emmanuel Levinas*. Springer.

Simmons, W. (1999). The third: Levinas's theoretical move from an-archical ethics to the realm of justice and politics. *Philosophy and Social Criticism*, 25, 83–104.

Smith, M. (2005). *Toward the outside: Concepts and themes in Emmanuel Levinas*. Pittsburgh: Duquesne University Press.

Tahmasebi-Birgani, V. (2014). *Emmanuel Levinas and the politics of non-violence*. Toronto: University of Toronto Press.

Žižek, S. (2006). Neighbors and other monsters: A plea for ethical violence. In S. Žižek, E. L. Santner, & K. Reinhard (Eds.), *The neighbor: Three inquiries in political theology* (pp. 134–190). Chicago: University of Chicago Press.