philosophical inquiries
www.philinq.it

philinq

philosophical inquiries

XI, 2
2023

Edizioni ETS

Essays published on “Philosophical Inquiries” are double-blind peer-reviewed.

Please visit www.philinq.it for submissions and guidelines.

six-monthly journal / periodico semestrale
Subscription (paper, individual): Italy € 50,00, Abroad € 80,00
Subscription (paper, institution): Italy € 60,00, Abroad € 100,00

Subscription fee payable via Bank transfer to
Edizioni ETS
Intesa San Paolo
IBAN IT 21 U 03069 14010 100000001781
BIC/SWIFT BCITITMM
reason: abbonamento “Philosophical Inquiries”

Registrazione presso il Tribunale di Pisa n. 1/13

Direttore responsabile: Alessandra Borghini

EDIZIONI ETS
Palazzo Roncioni - Lungarno Mediceo, 16, I-56127 Pisa
info@edizioniets.com
www.edizioniets.com
Finito di stampare nel mese di febbraio 2023

Distribuzione
Messaggerie Libri SPA
Sede legale: via G. Verdi 8 - 20090 Assago (MI)

Promozione
PDE PROMOZIONE SRL
via Zago 2/2 - 40128 Bologna

ISBN 978-884676845-2
ISSN 2281-8618

Table of Contents

Essays

Focus

philinq XI, 2-2023
ISSN (print) 2281-8618-ETS

Essays

Language versus reality
The case for phenomenology and the Deleuzian 'heresy'

Stathis Livadas

Abstract: This article is an inquiry into the relationship of language, as a phenomenon within the world, with the reality of the world as such and the ontological dimensions that underlie a conception of language in these terms. In doing this and in highlighting a kind of interiority of language with regard to reality naively thought, the author undertakes a discussion of the linguistic phenomenon in a broad phenomenological perspective, implying *ipso facto* a temporality factor, which except for an argumentation along this way deals also with the Deleuzian position on the matter in *The Logic of Sense*, as contrasted with the 'orthodox' or mainstream phenomenological view. A major place in the article has the argumentation about the deficiency of language in epistemological terms, more specifically in the face of certain phenomena associated with quantum mechanical situations.

Keywords: *Dasein*, individual, inner time, interiority, language, pre-predicative, quantum measurement, reality, sense, singularity-event, transcendental consciousness.

1. *Introduction*

If not for the second sentence to the title, the subject-matter of the article would be so wide-angled as would equivalently be, for instance, an article about the bounds of ontology in being within the world. Therefore limiting the discussion in terms of language as a phenomenon within the world, both in the 'orthodox' Husserlian view and the subsequent Heideggerian transcendental 'anthropocentric' position, served first of all to suspend the traditional rationalist approach of linguistics, which is to ignore the reality of language as a phenomenon in itself and consider it simply in what appears as an interjection between speaking and thinking by means of phonetic or written signs, i.e., a kind of codification mediating thinking with expression. In this sense the discussion draws to the source of linguistic phenomenon itself in a phenomenological perspective that would be the level of subjectivity put in absolute, non-reductionistic terms, well beyond the Chomskian attitude that smacks of a kind of subjectivist disposition and yet is being criticized for not fundamentally putting into question the deep structure of 'mental reality', in the sense that "linguistics

philinq XI, 1-2023, pp. 9-32
ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i2.417
Submitted: 9 June 2022
Accepted: 25 October 2023

should give us a picture of the 'mental reality' underlying language, which will then give us insight into the 'human essence' – into what distinguishes us from other life forms" (Grisham 1991: 38). On the other hand, choosing to talk about language vs. reality from a phenomenological viewpoint presented the chance to 'deconstruct' Deleuze's conception of a transcendental field, imbued with concepts of formal mathematics in a kind of putting the cart before the horses, as an attempt to found an ontology of beings in the world generative of the linguistic phenomenon by downplaying any kind of subjectivist concerns.

At the same time my critique of certain threads of Deleuze's thought, mainly in *The Logic of Sense*, may help clarify the idea of an 'interiority' of language that could be reducible to the 'interiority' of the subject, implying in turn a concept of inner, subjectively generated time. I refer, for instance, to Deleuze-Guattari's statement in *A thousand Plateaus Capitalism and Schizophrenia*:

> Not only are there as many statements as there are effectuations, but all of the statements are present in the effectuation of one among them, so that the line of variation is virtual, in other words, real without being actual, and consequently continuous regardless of the leaps the statement makes (Deleuze *et al.* 1969: 94).

In fact this is a kind of continuous variation that, Deleuze's eccentric, epistemically based metaphysics notwithstanding, leaves room for a possible interpretation in proper phenomenological sense:

> To place the statement in continuous variation is to send it through all the prosodic, semantic, syntactical, and phonological variables that can affect it in the shortest moment of time (the smallest interval). […] to content oneself with extracting a pseudoconstant of content, [which] is no better than extracting a pseudoconstant of expression. Placing-in-variation allows us to avoid these dangers, because it builds a continuum or medium without beginning or end. [..] A variable can be continuous over a portion of its trajectory, then leap or skip, without that affecting its continuous variation; what this does is impose an absent development as an 'alternative continuity' that is virtual yet real (Deleuze *et al.* 1969: 94-95).

Of course Deleuze had other inclinations than seeking a recourse to the human subjectivity in absolute terms for this kind of continuous variation. In *The Logic of Sense* events, even though are not confused with their spatiotemporal effectuation as states of things, are yet not thought but as essentially identical with meanings, the latter as what is inseparably the expressed or the expressible of a proposition and the attribution of a state of things (Deleuze 1990: 21-22). This is a position that would ineluctably end up in circularities or even conceptual overlappings by entering a notion of time that leads to a notion of temporal instants without 'thickness' conceivable

as mathematical points, to make room for a notion of events identifiable with meanings as 'incorporeal causes' extrinsic to linguistic propositions.

Leaving aside Deleuze's deviation from the broadly meant environment of phenomenological philosophy in which an essential part of the French philosophy of 20th century was nourished, my main focus in Sections 3 and 4 is to build up an argumentation for a phenomenologically founded view of language versus the world, one that would also reconcile certain aspects of the Husserlian and Heideggerian narratives on the matter. Primarily this would have to do, except for the inquiry into a pre-predicative level of discourse reducible to a priori forms of subjectivity, with the notion of inner time as 'coalescing' with the transcendence of subjectivity itself either in the sense of the Husserlian transcendental ego or in the sense of the Heideggerian *Dasein*. In these terms one may be able to found the 'interiority' of the linguistic phenomenon within the world on an 'interiority' prescribed by inner temporality as the essential mode of being of absolute subjectivity in taking also into account the intentional a priori modes of the latter. Accordingly one may provide a consistent account of meanings as ideal unities or species considered not as species of objects as such, i.e., in a material 'thingness' sense, but as species of intentional acts of thinking intimated in language use. Consequently, an 'interiority' of language in this sense as non-detachable from the world and yet not organic part of the world, could be reducible to the interiority of the subjective origin and attributable as a phenomenon, e.g., in virtue of Merleau Ponty's lived sense, to the embodied agency.

A major question dealt with in Section 5, namely the deficiency of language in capturing the being-in-the-world as unmediated by any constitutive-explicative faculties posed in principle a posteriori, seems to involve the epistemology of a situation in a purely worldly sense. What came out naturally as the field of preference to inquire into the relation of language, in the phenomenological perspective adopted throughout the text, with an epistemic situation is the field of quantum mechanics in which, for instance, the Heideggerian notion of 'being there' in actuality or the Husserlian notion of living present may possibly acquire a newly found relevance. This has especially to do with the quantum theory of measurement in which, more than probably anywhere else, the expressional capacity and the foundation of both formal and common language's 'interiority' *vis-à-vis* the world, are tested against the grindstone of physical reality.

Perhaps there is no better way to close the Introduction after the last epistemological prompt than quoting from B. D' Espagnat exactly as he wrote:

En conséquence la science ne se meut avec aisance que dans les domaines où le découpage – par la pensée – de la réalité en petits objets séparés est une opération féconde. Elle est

donc incapable de capter la vie dans ce que cette derniére a d' essentiel à savoir le fluide, le continu, et le mouvant.[1]

2. *The faultiness of the Deleuzian conception of transcendence as ontological – linguistic foundation*

Deleuze claimed in *The Logic of Sense* that dealing with sense not as a predicate or a property but as an event, more concretely (in the Deleuzian idiom) in terms of a nomadic or impersonal singularity, the discourse is no longer bearing the characteristics of a linguistic form as codifying the meaning-form, yet it is not about the formless but rather of the pure unformed (Deleuze 1969: 106-107).

However as I argue in the next, Deleuze's attempt to propose a foundation for the 'emergence' of meaning-forms irrespectively of subjectively founded a priori norms, is ineluctably bound to fail on the following grounds.

If we consider the genetic elements of a problem in general so that the category of sense replaces the category of truth, with 'truth' and 'falsity' based on the subjective and empirical level of knowledge, the relation that inheres between the problem and its conditions "defines sense as the truth of the problem as such" (Deleuze 1969: 121). Further if sense is intimately associated with the notion of event in Deleuzian metaphysics, and events are conceived as ideational singularities which communicate in one and the same Event that endlessly redistributes them in a way that their eternal truth extends indefinitely so long as they may emerge out of jets of singularities, (and thus justify their linguistic presence as infinitives), the whole point in Deleuze's argumentation against the inherence of subjectivity in the transcendental sphere ends up in a big circularity.

In the first place, if a problem is determined only by the singular points that express its conditions wherein singular points in the Deleuzian sense are meant as pre-individual, non-personal, a-conceptual and further as generators of a series of events in a determined direction up to the vicinity of another singularity, one may end up with a naive application of the mathematical notion of singularities. For instance, in the theory of differential equations the existence and distribution of singularities are relevant with the problematic field of solutions defined by a specific equation. It is common knowledge that singularities in the theory of differential equations and more generally in mathematical analysis are point-like 'deformations' of the

[1] "Consequently science does not move with ease but in the domains where the cutting – by the thought – of reality in small separate objects is a fertile operation. It is therefore incapable to capture life in what this latter has by necessity, that is, the fluid, the continuous, and the moving". See (D' Espagnat 2015: 121, auth. tr.).

mathematical continuum (think of the mathematical form of intuitive continuum), therefore they are subsequent to the continuous structure and by necessity cannot engender it. Furthermore, the kind of convergence or divergence in the vicinity of a singularity is implied by mathematical pathologies due to the structure of the real numbers as representing the continuum and not by the continuum itself in the sense of mathematized intuitive continuum.[2]

Time and again Deleuze applied the concept of Event (with a capital e) to describe its 'paradoxical instance' in terms of which all events are distributed and communicated in his own particular sense of nomadic distribution[3] that underlies the unique, aleatory and pre-objective being of the Event. Yet as with all other attempts throughout the history of ontological-metaphysical thought to dispose of the sense of being in absolute terms by means of an approach that would keep it totally unfettered from objectivist constraints, the concept of Event in the Deleuzian sense falls ultimately into the trap of having to account, in rejecting a reduction of a subjective type, for a being other to the Event in positioning the Event in a 'state' *ante* to that of denotation, manifestation, or signification something that naturally entails the pre-individuating, a-conceptual and non-personal character of the Event.

Evidently, phenomenologically thinking, this kind of actualization from the impersonal, pre-individual and a-thematical singularities to the individuated beings as persons would be accessible in no one's domain except for the domain of the latter beings as embodied carriers of a temporal consciousness and an I (*Ich*) for which the world has validity as a past that was, a present that is and a future that will be, all retrievable and presentable at once in the actual present as a streaming living experience. And there is clearly no possible way to have some kind of emission of singularities occurring on an unconscious surface by an immanent principle of auto-unification through a nomadic distribution without the presence of a subjectivity for which this state-of-affairs would be its own mode of being in the world as being-in-constituting thus and so. The negation of the latter supposition leads inevitably to the naiveté of a sort of objective or even physical realism evident from the way Deleuze relapses into mathematical conceptualizations to describe the (nomadic) distribution of sin-

[2] An idea of the intuitive continuum that comes easily to the mind is L.E.J. Brouwer's concept in relation with the primordial intuition of mathematics: i.e., the "substratum, divested of all quality, of any perception of change, a unity of continuity and discreteness, a possibility of thinking together several entities, connected by a 'between', which is never exhausted by the insertion of new entities" (van Dalen *et al.* 2002: 205).

[3] Nomadic distribution in the sense of 'emergence' of singularities is described by Deleuze as radically distinct from "fixed and sedentary distributions as conditions of the syntheses of consciousness"; see Deleuze 1969: 100-108.

gularities as 'crop ups' in a properly meant field of transcendence in which they arise as "topological events to which no direction is attached" and yet whose nature depends on objectively distinct instances (Deleuze 1969: 104-105).[4]

Consequently Deleuze entered, in a presumably epistemologically motivated context, into the same kind of circularities he accused the Kantian and the phenomenological tradition of having fallen into, namely by determining the transcendental field in the image of what it is supposed to ground. Of course by the latter allusion is meant the subjective sphere either in the Kantian concept of the synthetic unity of apperception or the Husserlian concept of the transcendental 'locus' of a priori intentionalities. In the same vein Deleuze criticized the Husserlian genesis for rendering the noematic nucleus of an object or event in the sense of a collection of attributes as a predicate and not as a verb, thereby insisting on the sedentary character of a concept and not on the 'kinematical' character of an event. Yet Deleuze has no other option to render the emergence of the impersonal and pre-individual singularities as intelligible out of the undifferentiated abyss except as 'realized' individual beings in allegoric mathematical forms. Even as the expressed world, i.e., the actualized world, is founded on the compossibility of different worlds conditioned on a mathematically inspired notion of convergence of the series of ordinary points around the vicinities of respective singularities, Deleuze once again slipped into the circular mode of relying on the founded to account for the founding. In this sense one would think of his statement that "the continuum of singularities is entirely distinct from the individuals which envelop it in variable and complementary degrees of clarity", (Deleuze 1969: 111), as nothing else than the ontological predominance of the actualized individuals over the pre-individual singularities in 'enveloping' them according to certain rules of convergence so as to be incarnated in a body or consist in single state out of a multiplicity of states, etc.

Even as Deleuze appealed to some kind of passive genesis to talk about a sense-generating world in which singularities-events are organized in circles of convergence, there is no reasonable ground to think of the pre-individual level of the transcendental field in any other way than in terms of actualization and individuation and consequently of expressibility involving by necessity a subjectively

[4] Deleuze, citing A. Lautman's *Le Problème du temps*, has argued, in an attempt to present the morphology of the solution of differential equations involving singular points as a substitutive ontology, that the nature of singularities as topological 'accidents' in the field of directions (relative to the form of a differential equation) is in a concrete sense defined by the form of the integral curves in their vicinity; (Deleuze, 1969: 344-345). However, as already alluded to, this is a kind of ersatz mathematized ontology that obviously downplays the implicit assumption of an *ad hoc* continuous substratum, in the sense of Brouwer's intuitive continuum, possibly reducible to the subjective modes by which one may have acquired this ingrained concept of continuum.

founded source. In these and these terms only, one can render a rational interpretation to Deleuze's claim that truth or falsity are transferred from the propositions to the conditions of a problem these propositions supposedly resolve, in a way that truth presupposes the sense attributed to the events. It would be paradoxical to construe out of these assumptions a transcendental field as opposed to the subjectivity of the person posited in absolute, non-mundane terms. In other words in the nebulous Deleuzian realm of singularities-events, the question of expression including a notion of expression in purely linguistic terms becomes equivalent to the question of actualization in the world involving individuals as presumably constituting (and not constituted) parts of the world.

Further and insofar as the actualization may be only conceivable in inner-worldly terms in the presence of a constituting consciousness and in the modes it is constituting, if there is a transcendental residuum in the linguistic constructs it would be rather found in the subjective sphere within-the-world. By this measure Deleuze's invocation of the univocity of Being, in referring both to what occurs and what is expressed, should be rather viewed as an attempt to do justice to a presumed transcendental element in the structure of language in denying at the same time any constitutional role to the subject. As it turns out a Univocal Being in the Deleuzian sense of "happening to things and inhering in language", would account for the interiority of language in the face of the exteriority of the world only by appealing to essential invocations of being of the kind found in traditional metaphysical arguments: the Being is neither active nor passive, it is extra-Being, "the minimum of Being common to the real, the possible, and the impossible" (Deleuze 1969: 180). And yet in the face of these allegations, indicative of an irresistible relapse to a kind of wide-angled subjectivism, Deleuze stated in the *Difference and Repetition* that the concepts of nature as concepts on an indefinite comprehension are found in the spirit that contemplates or observes and makes represent nature rather than in nature itself. On this account, nature itself is a self-opposing, alienated concept in the sense that the objects of nature do not possess and cannot recollect their proper moments. To cite an instance, rememoration, recognition and elaboration of memory in the natural repetition that necessarily refer to a *pour soi* of consciousness, as consciousness of knowing, is what is effectively lacking in a conceptualization of nature devoid of the constitutive capacity of an I. (Deleuze 1968: 14).

In the bottom line the Deleuzian interpretation of language, for instance, of the semantical content of the indeterminate infinitive in linguistic-grammatical form, as expressing the sense (or equivalently the event) in virtue of enveloping the 'internal' to the language time, seems ontologically lame insofar as the subjective constitutive factor is considered as little more than an exteriority to the event that is bound to express. In the face of it one may not bring up

an arbitrary transcendental scheme supposedly dissociated from subjectivist concerns to account for the 'interiority' of language in view of the exteriority of the world. For then, as will be further shown in the next sections, one would inevitably be dragged into unsubstantiated ontological assumptions or fall into the trap of reproducing circularities or yet succumb to both temptations.

3. *What should be the pre-predicative level in ontological and linguistic terms?*

If the methodological question concerning language as a tool of analysis has not been sufficiently addressed in the phenomenological literature, for the main reason that for a phenomenologist guided by the principle of eidetic intuition "once we have come into direct contact with the objects, the role of the concepts taken from ordinary language comes to an end",[5] it is still true that Husserl touched, although not extensively, on the issue in some places in *Ideas I* and in the *Logical Investigations*. In the latter and in a somehow indirect way Husserl focused on the concept of meaning-intention (*Bedeutungsintention*), as a signitive or symbolic intention to promote a view of knowledge as the fulfillment of such meaning-intentions without, as a matter of fact, taking into account in an explicit way the extent to which meaning-intentions are limited by the linguistic structure (Kung 1969: 331).

In yet another place, in *Formal and Transcendental Logic*, he sought, by means of the concept of predicative judgment (which in Husserl's view lies at the center of formal logic in its historical evolution), to reach by the syntactical deconstruction of a sentence of analytical discourse, the ultimate level judgment, foundational for all logical evidence, i.e., that of the ultimate pre-predicative experience reduced to the givenness of individual objects-in-person. The latter sentences in the general form of *S is p* would form the absolutely pre-logical level as it would be prior to any syntactical activity since, letting any kind of modalities out of question, more than any other linguistic form they are the outcome of a purely phenomenological evidence in dispensing with the necessity even of the objective existence of the syntactical subject *S*.

For Husserl this primordial predicative form must be the original noematic nucleus of all judgments, the structural foundation asserting something of something in the Aristotelian tradition of the declaratory proposition (ἀπόφανσις), from which all other derivative logical forms spring out as

[5] The quote is from an article of Husserl's disciple R. Ingarden published in 1919-1920 in a Polish journal. Exact citation may be found in Kung 1969, fn 3. In Kung 1969, Kung considered Ingarden's article as the only known exposition in some detail of the phenomenological method concerning the role of language.

syntactical superstructures (i.e., negation, conjunction, disjunction, modalization, quantification), by way of transformation or combination.[6] Consequently more than an assertoric proposition, Consequently more than an assertoric proposition, *S is p* becomes in the Husserlian sense a thetic proposition which by this virtue cedes the source of its originality, both in form and meaning, to the constitutive modes of a subject, and even more radically to their very origin. In this sense the question of establishing, by a genealogy of logic the pre-predicative level in both linguistic and ontological sense acquires a new content at odds with attempts to attribute it, like Deleuze and various metaphysical philosophers did before him, to some transcendental field 'extrinsic' to subjectivity. In these terms the reduction of the self-evidence of judgments in the objectual self-evidence may entail the question of whether the pre-predicative level, reached by syntactical regression in the first place, would indeed lead to the pre-logical level and, still more important, will raise the question of the nature of the procedures required for leading methodologically *a tergo* to the pre-predicative universe. The apparent methodological steps leading back from derivative to original judgments till the ultimate level of irreducible substrates, evident in the intentional experience of 'thingness' or completely abstract individuals, are syntactical operations and nominalizations of such operations by which we reach an ultimate level unfettered by any syntactical concerns and thus foundational for all logical evidence. This would presumably be the level of ultimate pre-predicative experience, understood as the givenness of individual objects in the Aristotelian sense of a categorially irreducible 'general-something' (τόδε τι). These primordial, non-analytically reducible objects-individuals would form the absolutely pre-syntactical and in a sense pre-logical level, i.e., the one prior to any syntactical activity. In this sense one may pass from the domain of logos understood as a correlate of meaningful acts of expression and ideal significations to the domain of logos as intentional correlate of acts oriented to the 'lowest' level of intentional apprehension, i.e., prior and foreign to all logical and consequently syntactical activity.

Yet if by eliminating all acts of syntactical construction we may be brought back from an upper substrate to the immediately lower one and this way to the ultimate substrates given in the sheer experience of individuals as such, how could it be possible, one may ask for example, to interpret the supposedly syntactical reduction from the mathematical cardinality of

[6] Contrary to the traditional positing of *S is p* as the original form of categorical judgment admitting of two juxtaposed co-original forms, namely, the positive and the negative judgment, Husserl favored the original predication *S is p* as a single-layered (*einsichtig*) one, to the extent that it implies a "nominal position understood as a fundamental position", i.e., the positing of a substrate or object as a subject attributed with a predicate *p*, which by this positing alone implies a sense of subjective constitution.

sets in general to the set of real numbers, 'from the latter to that of rational numbers, and, in like fashion, to natural integers, then to singular integers understood as a multiplicity of units, and finally, to the individual objects from which they were drawn by formalization'? (Pradelle 2021: 61-62).

If the methodological steps in passing from the level of more complex judgments to the lowest level, namely that of the evidence of irreducible substrates, are thought solely in terms of syntactical deconstruction one may hardly account in this context alone for the reduction from the 'mathematical cardinality of sets' in general to that of the set of real numbers, from the latter to that of the set of rational numbers and so on. For anyone knowledgeable with the current and last century's developments in the foundations of mathematics a key issue brought up in the epistemology of mathematics, in fact in the ontology of mathematics, is the question of the deficiency of syntactical means to 'capture' not strictly finitistic mathematical concepts, a case highlighted by Gödel's incompleteness results and the still pending decidability question of the well-known *Continuum Hypothesis* involving the cardinality of the mathematical continuum. (See, for instance, Fefeman 1999; Livadas 2019, 2020).

In *the Experience and Judgment* Husserl characterized the colligation of objects A, B, C,..., syntactically nominalized as the conjunction of A, B, C, ..., in the form of the set {A, B, C, ...}, as essentially founded not on material elements nor on the essence of things themselves insofar as their essence is taken into consideration only as it makes differentiation possible (Husserl 1964: 188-189). Instead, to make a collection of objects (e.g., a set of objects or a class of sets of objects) a thematic object in actual presence, an act of a higher order level is required, one of productive spontaneity rather than one of passive receptivity. In a showcase of the insufficiency of syntactical means to capture the conception of a whole, irrespectively of the cardinality and the essence of its constituting elements, as a completed unity in actual presentation Husserl appealed to what he termed a retrospective apprehension (*rückgreifendes Erfassen*). Perhaps not unexpectedly, given Husserl's constant preoccupation in his post-*Logical Investigations* years with the origin of transcendence within immanence, this was meant as an act of thematization of a collectivity of objects by the constituting (transcendental) ego, into an identifiable and re-identifiable object-meaning possibly posited as a substrate of judgments in general and, in particular, of formal-mathematical propositions (Husserl 1964: 246-247). This kind of constituting activity was meant in fact as a unity-constituting and consequently a meaning-founding act of the transcendental ego as ego-in-act. Consequently it might possibly lay the ground to justify the transcendental element found in the notion of the 'interiority' of language as pertaining to the being-in-constituting of the ego itself,

implying as a matter of fact its mode of being as temporal. Obviously it is in this sense that must be read Husserl's conditioning of the logical requirement of individuality on the unique (inner) time, that is, of the "requirement of an object as an identical substrate of predicates and of objective truths (subject to the principle of noncontradiction)" and further of the idea of a whole of interconnected possibilities (Husserl 1964: 355-356). Put succinctly:

> Now every intuition we have, every phenomenological perception, memory, etc., every judgment, every statement, sense, conscious intention is absolute consciousness, and all this in the unity, that belongs to these experiences. Naturally we have there to turn back to the ultimate flux of time and we have to think of all unities drawn back to their last and fundamental multiplicities (Husserl 2013: 139).

This is of course a view tied to a conception of objectivities as ontologically dependent on the absoluteness of consciousness in the sense that certain a priori features of absolute consciousness as the unity and the interconnection of conscious experiences (as immanent appearances) cannot be attributed to physical laws. This leads as a consequence to an idea of 'indestructible' objectivity apt for application by means of a meaningful linguistic environment to the extent that the 'lowest' grounds of scientific objectivity are due to invariances inherently associated with ultimate, non-eliminable forms of objectivity, beyond any notion of beginning and cessation (Husserl 2013: 151).

As a matter of fact for both Husserl and Heidegger, even as Heidegger was gradually distancing himself from Husserlian phenomenology and its promulgated transcendence within the immanence of consciousness, the 'interiority' of language implied by the founding unity of any meaningful discourse would be ultimately associated with an absoluteness established in subjective terms and by implication hinged on inner temporality.

In turn Husserl's radical reduction to the transcendence of the ego, to the extent that the regression from the logical structures of signification involves the noetic[7] and noematic structures present in the 'lower' layers of intentional apprehension, brings into the foreground questions that touch on the transcendence as founded on the absoluteness of subjectivity itself. This said, if the noetic-noematic level of intentional apprehension is meant as preceding meaningful forms of linguistic expression one may be rightfully reserved as to the possibility of properly founding the presumably pre-objective

[7] A noematic object is an object said to be constituted by certain a priori modes as a well-defined object (an object as meant), immanent to the temporal flux of a subject's consciousness. In contrast to noematic objects, noetic objects described as moments of hyletic-noetic perception can be only thought of in terms of evident 'givennesses' of the a priori orientation of intentionality. More in Husserl's *Ideas I*: Husserl 1976: 229-232.

character of noetic enactment within the sensuous field of experience. This seems an open question related more generally with the Husserlian conception of the pre-objective character of intentionality in view of the necessity to appeal to the reflection itself, a necessarily objectifying act, to be conscious of any intentional act. Husserl has in fact left, as it happens also with the 'ontology' of transcendental ego, the question of the objectivity of intentional acts as such in suspense. In *Phenomenological Investigations* (Suppl. volume, part II), for instance, he has clearly stated that each act of the objectifying cogito oriented to an object, whose being is posed as thematic, is an actual intention that is objective. This also applies to the special case of meaning-intentions for which, in Husserl's words, we do not know yet whether one can have non-objectifying acts as meaning-giving ones (Husserl 2005: 200).[8]

In view of the above we may have to regress to a 'world only for me' in order to reach the pre-predicative and therefore pre-logical level of experience, by abstracting from the limited intersubjective validity of the language we speak[9] and further by going back from the founded experiences, e.g., cultural or epistemic objects, to the simplest sensually accessible ones. Could there be, in such terms, a residuum of the world reducible to sensuous perception alone, a world of exclusively sensuous substrates, of primary substances, and of bodies as given in external experience allowing to establish lowest-level, pre-predicative judgments ultimately appealing to individuals as irreducible, sensuous substrates given in the simplest form of predication *S is p*? One has serious reasons to doubt, insofar as the kernel of lowest level judgments, the non-analytically reducible τόδε τι, supposedly deprived, in *Formal and Transcendental Logic*, even of a temporal form and considered as just an intentional correlate, has relegated its ontological legitimacy from the world of external experience to the experiencing subject as temporally constituting in absolute terms.

4. *In what terms Heidegger and Husserl shape the discussion on the relation of language to the world?*

If there is a common thread to judge Husserl's and Heidegger's treatment of the ontological foundation of language it is primarily the need to account for the role of language *vis-à-vis* the world with all that this position implies in terms of subjectivity, temporalness and straightforward representation,

[8] This kind of ambivalence regarding a presumably non-objective character of intentional acts and the ensuing circularities may be found in various places in Husserlian texts, e.g., in Husserl 2006: 113), (Husserl 1973: 543, 550, Husserl 1968: 353, 423.

[9] See Pradelle's arguments in Pradelle 2021: 68.

according to which linguistic structures correspond to phenomenal features. On this account the possibility of application of linguistic forms on the basis of an 'empty' content in contradistinction with phenomenal 'fulfillment' awareness underscores the non-existence of an isomorphic mapping, to use mathematical parlance, between linguistic forms and features of the world, something that was a common preoccupation for both, especially concerning the routine language use in Heidegger and the 'puzzle' of symbolic thinking in Husserl. In the post-*Logical Investigations* years Husserl faced the challenge of the aforementioned 'puzzle' by employing, in the *Formal and Transcendental Logic*, the concept of 'anything-whatsoever' (*Etwas überhaupt*) in a formal-ontological sense applicable, primarily, in propositions involving formal-mathematical individuals, corresponding to 'empty' intentional substrates devoid of any material content whatsoever (Husserl 1974: 77-78).[10] Given the capital importance they both attached to the role of temporality as co-constituting a non-reductive subjective foundation of being in the world this was to be reflected in what would determine language as human activity within-the-world.

More concretely for Husserl:

> Time consciousness is the original seat of the constitution of the unity of identity in general.[..] The result of temporal constitution is only a universal form of order of succession and a form of co-existence of all immanent data. But form is nothing without content. Thus the syntheses which produce the unity of a field of sense are already, so to speak, a higher level of constitutive activity (Husserl 1964: 73).

In these terms the temporal form is not only a form of individuals, to the extent that we may talk about enduring individuals, but may further have the function of uniting individuals in a unity of connection (Husserl 1964: 158). It is noteworthy that Husserl's conception of logical-linguistic activity in subjective-temporal terms underwent a gradual evolution virtually from the time of *Logical Investigations* onwards, wherein the turn to a transcendental-subjective foundation was becoming more and more evident. In *Logical Investigations II*, for instance, meaning is characterized as the ideal species of intentional acts pertaining to non-separable 'qualitative' and 'material' parts as unity, and further meanings as ideal unities or species are considered not as species of objects as such but as species of intentional acts of thinking intimated in language use (Husserl 1984: 122-123, 308-309). Consequently for Husserl the unity of perception of a plurality of individuals, a unity on the basis of

[10] The distinction between 'empty' substrates and associated syntactical objectivities and 'thingness' substrates and associated 'materially filled' syntactical objectivities corresponding to material objects is also found in *Ideas I*; Husserl 1976: 27-28.

a connecting temporal form, to the extent that temporality has been 'interiorized' in transcendental reduction, has served as the foundation of formal-ontological unity in the sense of a special kind of constituted unity that provides the basis for special relations, namely, the formal relations appealing to empty-of-content 'general-somethings', and further to the concept of language itself.

Heidegger, on the other hand, had associated a notion of temporality with language in terms of the demonstrative function of articulacy in the sense of the latter as participatory communication (*Mitteilung*) in being-in-the-world. In *Being and Time* he pointed to the temporality of discourse meant as 'interiority' that should be neither confused with a vulgar sense of temporality insofar as language speaks about temporal processes in the various tenses employed a propos, nor with the fact that talking occurs in 'psychical time'. Heidegger's concept of the temporality of language is plainly stated as following:

> Discourse is in itself temporal, since all speaking about .., of .., or to … is grounded in the ecstatic unity of temporality. The kinds of action are rooted in the primordial temporality of taking care of things, whether it is related to things within time or not. With the help of the vulgar and traditional concept of time which linguistics is forced to make use of, the problem of the existential and temporal structure of the kinds of action cannot even be formulated (Heidegger 1967: 320).

While leaving, for instance, the notion of the present in ambiguity[11] in that the now-saying *Dasein* 'understands itself in terms of what it is available in the world', Heidegger outlined in *The Concept of Time* his commitment to the non-reductive character of *Dasein*'s temporal being in the world reflected in a temporal conception of language as a basic mode of being-in-the-world. In these terms, prior to the way language expresses time thematically, comes the more fundamental question of how the temporalness (*Zeitlichsein*) of being-in shows up in language, in which case a theory of tenses founded on the temporal being of being-in of *Dasein* (in the particular Heideggerian sense of self-alienation) would be the plausible way to look back to the basic foundations of traditional grammar. One may think, a propos, of futuralness as expectant temporalness becoming everydayness 'to the extent that being-in succumbs to the world' (Heidegger 2004: 63-64).[12]

[11] I have in mind the ambiguity concerning, on the one hand, that which is the present in the surrounding world (*die Präsenz*) and, on the other, the present now as lived experience of *Dasein* itself (*das Präsens*). See Heidegger 2004: 63.

[12] The Heideggerian notion of language as being itself temporal does not contravene the deposition and 'exact' reactivation of the formal signs model, itself 'approximative or schematic in character', allegedly implying an essentially atemporal relationship between expressions and their sense-genetic origins (Inkpin 2016: 80). Heidegger's implication of time in terms of linguistic activity is of a

On a shared phenomenological background Husserl's view of the words is that they are not just signs, bearers of a semantic content, but 'vectors of meaning in the sense of acts of intending' in a way that "the verbal and semantic consciousness are not juxtaposed to one another, disjointed, but rather, make up a unity of consciousness in which the double unity of word and sense [*Wort und Sinn*] is constituted".[13] Husserl moreover claimed that the intentionality unifying the words themselves and the sense, the living experience of the word and the thinking, has the character of patent intentionality which in contradistinction to the latent intentionality presupposes the active presence of pure ego (Husserl 1974: 366).

As known, the origin of the Husserlian pure ego was never clarified whether it might be derivable by an *in rem* concern over the subjective origin of the synthetic unity of the world in the Kantian tradition or by the purely logical necessity of breaking off the interminable chain of constitutive causes. Consequently even as the concept of the pure ego is regarded the 'black hole' of the Husserlian transcendental reduction, yet this kind of radical reduction unifying word and sense would by all accounts mean that if there is a transcendental factor in the 'interiority' of language in relation to the phenomena of the world then this should be associated with the kind of transcendence found in the 'interiority' of the subject itself with all that implies with respect to a subject's a priori constitutive modes.

I draw attention here that in a broadly conceived converging perspective with the phenomenological attitude a conception of language 'without recourse to an ideal of full, nontemporal determinacy' makes Wittgenstein's and Merleau Ponty's views compatible on the matter insofar as Merleau Ponty rejected any ideal of full determinacy in considering linguistic meaning as characterized by constitutively indefinite horizons in the process of formation, while Wittgenstein was essentially of the same view to the extent that 'the commitment to full determinacy implicit in his earlier calculus model of language leads to incoherence' (Inkpin 2016: 220). Wittgenstein's calculus model of language in his *Philosophical Investigations* failed on the grounds that the regress-of-rules argument would imply that a calculus-underpinned language lacks of a proper foundation as it renders inconceivable the ideal of full determinacy insofar as it generates a non-terminable regress of meaning-

deeper genetic origin inhering in the essence of being of *Dasein* as temporal and in the ecstatic unity of temporality. Further, it is not true that Heidegger had generally (beyond SZ) 'nothing specific to say about the temporality of language (either *Rede* or *Sprache*) as such', as claimed by Inkpin in (Inkpin 2016; note 29: 325). In fact Heidegger does so explicitly, though not extensively, in Heidegger 2004: 74; 63.

[13] See Vandevelde 2021: 199-200, 203-204, 209.

attributing rules and hence incoherence. Concerning, however, Wittgenstein and Heidegger, while both conceive language in a purposive perspective inasmuch as Wittgenstein's intrinsic link between the use of signs and forms of practice may be thought to enlarge the context of Heidegger's instrumental relations involving the use of linguistic signs, it is still phenomenologically unfounded to draw analogies, as Inkpin does,[14] between Wittgenstein's association of linguistic signs and forms of practice with Heidegger's derivation of the significance of words from *Dasein*'s circumspective setting-out. Indicative of the vagueness of the demarcation line between the transcendental and the mundane spheres, this means that *Dasein*'s circumspective setting-out may have a transcendental origin founded in the mode of being of *Dasein* itself well beyond Wittgenstein's mundane interpretative undertaking on the issue.

After all language for Heidegger, as a primary ontological mode of the public realm, in all its phenomenal reality must be referred back to *Dasein* as a way of *Dasein*'s being and its modes of being. On these grounds *Dasein*'s predicative awareness, characterizing Heidegger's conception of language as the modification from purposive to an objective properties-based individuation of entities, may be neither conceived through an ontologically separated, ego-independent 'inside-outside' of language nor through language as an autonomous or abstract entity that comes into contact with the world only accidentally (Inkpin 2016: 224). Furthermore Heidegger's conception of language, more specifically, the non-inferential grasp of the features of the world in the disclosing function of linguistic signs in the sense that "all disclosure of the world is embedded or founded in pre-predicative equipmental or purposive awareness, a view that […] extends to the use of language" (Inkpin 2016: 227), points to the founding role of the pre-predicative level in terms of language formation in a way reminiscent of Husserl's invocation of a pre-predicative level to accede to the most fundamental level of logical-linguistic activity as discussed in Section 3.

If along these tracks one may vindicate a view of language that is more than an intellectually structured complete and rationally functioning system of signs, in which the pre-predicative level of linguistic experience can be 'interiorized' as founded on a special kind of 'interiority' of the subject, one may get a linguistic activity which even as a phenomenon referred to and conditioned by being-in-the-world it is still in excess of pure mundanity. This means that, far from any ad hoc *mélange* of metaphysical and

[14] Inkpin is oriented to a conception of language, in the sense of a so-called minimalist phenomenology of language, that is more close to a version of cognitive theory than to a transcendental phenomenology properly meant. Consequently he is bound, contrary to the Husserlian or Heideggerian views, to treat the question of the phenomenology of language in essentially mundane, objectivist terms. See Inkpin 2016: Ch. 10.

epistemological notions, seemingly Deleuze's way in the *The Logic of Sense*, language may be 'interiorized' as inalienably associated with the mode of being in absolute terms of an embodied consciousness in whatever particular denomination this latter may be found in the continental philosophy literature.

5. *The deficiency of language in the epistemology of the situation*

If language as a phenomenon within-the-world has an 'interiority', possibly thought of as the residuum left over after the elimination of all acts and apprehensions taking place in the world as expressible in standard linguistic forms, and if this 'interiority' may be reduced to the 'interiority' of the self as the absolute subjectivity factor, then the epistemology of a concrete quantum mechanical situation may prove a terrain of predilection to provide a convincing evidence for such claim. Especially if this situation tests *in extremis* the capacity of language to express by its linguistic means the process of being in being-objectified, in case we do not take recourse, for example, to the Deleuzian eccentricities of seeking the origin of sense in the so-called nomadic or impersonal singularities that refer in turn to an allegedly pure unformed being in banishing any kind of hetero-determination. By the same rationale one might inquire about the capacity of language to represent such categorial objects of mathematics, as the infinite sets or the formal individuals, in the Husserlian sense of formal-ontological objects, as consummate objects in terms of a subjective constituting activity. Naturally this kind of discussion may involve at some point the clarification of the role of subjectivity as transcendence and the grounding of its 'being there' in the actuality of the world together with the consequent involvement of inner temporality. The upshot of this inquiry reaching to phenomenological concerns about deep language structure is that the involvement of temporality, in the sense alluded to already, brings out deeper questions that stand the core matter of the phenomenological inquiry itself.

According to the Husserlian narrative, the unity of temporality as an objectivity leaves *in rem* an 'ontological' vacuum between the non-reflective, pure ego itself and its enactment in the present 'now and here', whereas for Heidegger the ecstatic unity of temporality, that is, the unity of the 'alienation-of-itself' in the raptures of past, present, future is the condition of the possibility that an existent can be as its 'there' (Heidegger 1967: 321 in: Livadas 2022: 2-3).

A sense of being as 'being there', implying a sort of inner temporality on the part of the 'questioning entity', i.e. the questioning subjectivity, may 'naively' and in indirect fashion vindicate itself in the way the separation between conceptual and factual in general may be considered a fuzzy one. For instance, in Quine's *Two dogmas of empiricism* the ontological core of our field

of knowledge is underdetermined by the boundary conditions of our experience insofar as properties can never be sufficient enough toward a complete and each time unique description of objects themselves, an almost obvious truth in quantum theory. Moreover objects in general, in particular quantum ones, except for material objects may also be considered as objects (or relations) re-presentable in abstraction in the context of a formal-mathematical (meta)theory, consequently as constrained not only by their status as physical objects but also as formal-mathematical ones. As already discussed in Section 3 the latter ones in virtue of formal-ontological objects imply, at least in the Husserlian narrative, the constitutive capacities of a transcendental subjectivity. On this account, the 'questioning entity', which can confirm its 'being there' as an unambiguous evidence in the present now and in the modes it constitutes objectivity, can shape an ontology of the situation that may conflate with epistemological concerns both in the level of 'observation' and, to the extent that mathematics as a formal syntax bestowed with modes of meaning-in-the-world is a highly specialized linguistic activity, also that of language.

In broad terms the question of being as reformulated into a question of a subjectively founded 'being-there-in-actuality', that is, being originally in the living present, may ground each subject's temporal particularity and establish the foundation of each individuality in the world, independently of context, as identically and invariably the same for that matter. In that case a sense of individuality in purely subjective terms and in the specific 'being there' of absolute subjective origin would be the ultimate foundation of the definiteness of a situation/state-of-affairs in the actual present irrespectively of whether we are talking, for instance, about the disentanglement of a quantum state-of-affairs upon 'observation' or about the constitution of an infinite formal-mathematical object out of an ideally infinitely proceeding mental construction. In both cases one can make possible a formal discourse about phenomena-in-the-world turned to meaningful linguistic objects out of subjectively founded processes that are yet non-eliminable by purely linguistic means. In this respect a subjectivity grounded in absolute terms making itself an unambiguous presence in actuality may pertain to the 'being there' in epistemological sense. Put in Husserlian terms, the transcendental ego by its very enactment in the living present, which is naturally not to be meant as a common sense self-awakening, nullifies the ontological vacuity between consciousness as passive receptivity (reflected upon) and consciousness as consciousness-of (reflecting on). This transcendentally founded act may be epistemologically read, in terms of quantum measurement, as the possibility of identification of the quantum state registered by a detector with the consciousness of the same state by a time-constituting transcendental ego. By this token one may view through another

angle, on the one hand, the conceptual ambivalences concerning the objectivity of the state vector in the case of wave packet reduction, and on the other, the possibility of idealist interpretations associated with Bohr's assumption of the non-objectivity of the state vector (D' Espagnat 1999: 90-91, 253, 259).

In such terms a notion of the living present meant as the way of 'being there' of the transcendental subject in the particular situation might prove worthy of further discussion in epistemological terms as it bears on the way a process of being-in-constituting 'transforms' into a solidified objectivity transformable on an intersubjective basis into a linguistic object of a meaningful discourse. As stated before, the quantum-theoretical context as most inherently related with the subjective modes of 'observation' in being-in and facing-up to the world proves to be a field of preference to discuss the foundation and the bounds of linguistic activity with regard to phenomena within the world.

In these terms if one forms an idea of the living present as the undoubted self-confirmation of each subject's mode of existence in the world, one by which he has the sole and unique mode of accessibility to the world of phenomena including his own self, we may well come to conclude that the observational language of quantum mechanics may be only interpreted classically for it involves the self-enactment of the interacting *I* (*Ich*) in each living present in terms of the triangle conscious subject – measuring apparatus – quantum-state-of-affairs, expressible only in the state of objectification. Yet there seems to be more at play here than just an observational-theoretical division between classical terms as representing 'observational' ones and quantum terms as representing 'theoretical' ones, for which the orthodox Copenhagen interpretation appealed to the 'extra-physical' notion of quantum state collapse. To the extent that the rationality of nature makes it generally possible to have a mathematical physical theory in the formal terms of which one may account for the past and contemplate for the future events, the residue emerging de facto between the unitary evolution of a quantum state-of-affairs and the classically interpreted language of post-measurement outcomes is bound to re-appear in another form in the structure of the linguistic metatheory as a concrete demonstration of the non-eliminable 'interiority' of language itself in the face of certain phenomena-within-the-world. Rather than having to rely on realist accounts or contextual theories of meaning, in fact unable to provide a satisfactory account of the approximations involved in the transition from the quantum mechanical to the classical level, and of course having much less in common with Deleuze's idiosyncratic metaphysics in *The Logic of Sense*, a properly meant phenomenological account of the 'residue' in the quantum 'observation', turned into a linguistic 'approximation', would prove a luring interpretational means especially in view of its appeal to the absoluteness of the living present as mode of being

of the subjective factor. However, as I will claim in the next, this is exactly what makes language forever missing nature, the latter as authentically being itself.

It is known that Bohr went so far as to assert that we have no other means of understanding quantum mechanics other than the classical ones, in the corresponding linguistic norms based on a self-standing objectivist interpretation of nature. In this view "the appropriate physical interpretation of the symbolic quantum-mechanical formalism amounts only to predictions, of determinate or statistical character, pertaining to individual phenomena appearing under conditions defined by classical physical concepts" (Bohr 1949: 210-211, 238). It happens that von Neumann's reduction postulate, being a high profile case of the relevant argumentation, has rendered impossible at least in the ontological level of a quantum measurement to account for the definiteness of post-measurement values of quantum observables without the implicit acceptance of the consciousness of a participating subject.[15] For such subject a notion of a self-constituting inner time in terms of which he must 'act', in the absence of any sort of reflection (including self-reflection), should have to be prior established. It follows that the acting subject's participation in the measurement process cannot be subsumed to a kind of physical reductionism by "evoking some physical event that occurs in the brain of the observer at the end of a measuring interaction. For such event would remain 'inside the (quantum) calculation' and would therefore do nothing to break the chain of entanglements and superpositions" (Bitbol 2021: 571). Then if one does not concede to some kind of 'ghostly' property of consciousness which can make possible a collapse of quantum states and the attainment of the 'linguistic level' of post-measurement values, the reduction to a constituting subjectivity conceived in absolute terms seems to be the plausible way between the Scylla of physicalistic reductionism and the Charybdis of eccentric metaphysics. Bitbol, evoking von Neumann's use of the quasi-Husserlian expression 'abstract ego', has aptly referred a propos to von Neumann's view in that "the

[15] Given that due to its philosophical orientation the present article cannot enter into the technical details of the issues in quantum theory involved, the author suggests for those interested for a further reading, among many other sources (Boge 2018; D' Espagnat 1999; von Neumann, 1955). Concerning von Neumann's reduction (or projection) postulate, which essentially amounts to the supposition that consciousness is able to modify physical states by collapsing them from superpositions of states to sharp values, there have been various alternative interpretations, among them Feyerabend's in a 1957 paper, dispensing with the idea of a quantum unitary evolution collapse on grounds contrary to positivist ones. Yet the efforts to provide a link on statistical grounds between the uninterpreted formalism of unitary evolution representing a quantum state-of-affairs, as being in itself an 'unknown' process, and the classically interpreted language of post-measurement outcomes, have shortcomings on their own as the relation between observers and macroscopic measurement devices includes more data than is typically appreciated, while leaving out of account decoherence effects.

divide between the observer and the observed system can be moved back further and further until nothing (not even a brain, not even a ghostly soul) is left on the observer's side. It can be moved until the observer is represented only by her 'abstract ego', namely by a pure knower unknowable to itself, whereas all the rest is treated as a global (quantum) system" (Bitbol 2021: 572).

In other words in order to avoid the trap of physical reductionism one may with good reason make room for a consciousness in absolute temporal terms whose act of self-constituting would be 'inaccessible' to its reflecting self for then it would be part of the global physicalistic quantum system and thus inappropriate to account for the residuum lying between the uninterpreted formalism of the unitary evolution of superposed quantum states and the classically interpreted language of post-measurement outcomes. In a certain sense one comes across a persisting conundrum of phenomenological reductionism, namely the way to found a temporality-constituting consciousness that would not be identically consciousness of itself and consequently asking for a purely subjective origin of its own self in an interminable recurrence. Which is to say, what lies ahead is the way to found a kind of ever-in-act 'substrate' of consciousness that would always 'elude' reflection and such that it would also account for the interiority of language in resolving the 'being-in-the-flow' of the world as being 'already there' and in consummate objectivity. This kind of experience of the present, attributed to von Neumann's subjectivist account of his reduction postulate, in Bitbol 2021, as essentially a sort of constant self-awakening of the subject and a means to 'fill in' the chasm between living as original presence and thinking about living as original presence is part and parcel of the phenomenological discourse in both Husserl's and Heidegger's respective narratives.

Appealing to the living present in the terms discussed above, i.e., as the possible means of 'appearance' of the ego within the world, may offer a clue as to the possibility of eliminating the residue between acting-in-actuality and reflecting upon acting-in-actuality.[16] However the kind of ontological vacuity, re-presented as a 'residue' in quantum terms between the 'being-in-entanglement' of a quantum state and its registration as post-measurement valuation, turned into immanent vacuity by transcendental-subjective considerations, may eventually prove non-eliminable due to the exclusively objective means available to put it into evidence. And by this measure the linguistic means available, to the extent that language amounts to a kind of normativity with regard to what has already come ontologically to 'be there',

[16] See, e.g., Husserl's references in *Späte Texte über Zeitkonstitution* to the way the pure ego, as abstractness, becomes concreteness through the 'content' of the streaming present (Hussel 2006: 29, 53). Also Heidegger's reference to the being-there of *Dasein* as what it is in the initial givenesses now and soon to come; among other places, in Heidegger 1988: 24, 28, 65-67.

are most probably bound to leave the 'interiority' of language, in the phenomenological sense bestowed in this article, untouched and the conjecture of whether reality will forever elude language *essentialiter* unanswerable.

Stathis Livadas
Independent Scholar
livadasstathis@gmail.com

References

Bitbol, Michel, 2021, "Is the life-world reduction sufficient in quantum physics?" in *Continental Philosophy Review*, 54: 563-580.

Boge, Florian, 2018, *Quantum Mechanics Between Ontology and Epistemology*, Springer, Cham, Switz.

Bohr, Niels, 1949, "Discussion with Einstein on Epistemological Problems in Atomic Physics", in: Schilpp, Paul, ed., *The Library of the Living Philosophers. Albert Einstein: Philosopher-Scientist*, MJF Books, New York: 200-241.

Deleuze, Gilles, 1968, *Différence et Répétition*, PUF; Paris; Engl. tr. by Paul Patton 1994, *Difference and Repetition*, Columbia University Press, New York.

—, 1969, *Logique du Sens*, Minuit, Paris; Eng. tr. by Mark Lester, Charles Stivale 1990, *The Logic of Sense*, Columbia Univ. Press, New York.

—, Guattari Felix, 1980, *Mille plateaux: Capitalisme et schizophrénie*, Minuit, Paris; Eng. tr. by Brian Massumi 1969, *A thousand Plateaus Capitalism and Schizophrenia*, University of Minnesota Press, Minneapolis.

D' Espagnat, Bernard, 1999, *Conceptual Foundations of Quantum Mechanics*, Perseus Books, Reading, Mass.

—, Bernard, 2015, *A la recherche du réel*, Dunod, Paris.

Fefeman, Solomon, 1999, "Does mathematics need new axioms?", in *American Mathematical Monthly*, 106: 99-111.

Grisham, Therese, 1991, "Linguistics as an Indiscipline: Deleuze and Guattari's Pragmatics", in *SubStance*, 20, 3, 66: 36-54.

Heidegger, Martin, 1967, *Sein und Zeit*, M. Niemeyer Verlag, Tübingen; Eng. tr. by Joan Stambaugh 1996, *Being and Time*, State University of New York Press, Albany.

—, 1988, *Ontologie- Hermeneutik der Faktizität*, Klostermann, Frankfurt; Eng. tr. by John van Buren 1999, *Ontology – The Hermeneutics of Facticity*, Indiana University Press, Bloomington.

—, 2004, *Der Begriff der Zeit*, Klostermann, Frankfurt; Eng. tr. by Ingo Farin 2011, *The Concept of Time*, Continuum, London.

Husserl, Edmund, 1964, *Erfahrung und Urteil*, Claassen Verlag, Hamburg; Eng. tr. by Churchill, James, Americs, Karl 1973, *Experience and Judgment*, Routledge & Kegan P., London.

—, 1968, *Phänomenologische Psychologie: Vorlesungen Sommersemester 1925*, Hua IX, Springer, Dordrecht.

—, 1973, *Zur Phänomenologie der Intersubjektivität. Texte aus dem Nachlass. Dritter Teil*, Hua XIV, M. Nijhoff, Den Haag.

—, 1974, *Formale und Transendentale Logik*, Hua XVII, M. Nijhoff, Den Haag; Eng. tr. by Dorion Cairns 1969, *Formal and Transendental Logic*, M. Nijhoff, The Hague.

—, 1976, *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie*, Erstes Buch, Hua III/I, Den Haag, K. Schuhmann; Eng. tr. by F. Kersten 1983, *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy: First Book*, M. Nijhoff, The Hague.

—, 1984, *Logischen Untersuchungen*, Hua XIX/1 & XIX/2, M. Nijhoff, Den Haag; Eng. tr. by Findlay 2001, *Logical Investigations Volume II,* Routledge, New York.

—, 2005, *Logische Untersuchungen. Ergänzungsband Zweiter Teil*, Hua XX/2, Springer, Dordrecht.

—, 2006, *Späte Texte über Zeitkonstitution, Die C-Manuscripte, Hua Materialien Band VIII*, Springer, Dordrecht.

—, 2013, *Grenzprobleme der Phänomenologie*, Texte aus dem Nachlass (1908-1937), Springer, Dordrecht.

Inkpin, Andrew, 2016, *Disclosing the World*, MIT Press, Cambridge Mass.

Kung, Guido, 1969, "The Role of Language in Phenomenological Analysis", in *American Philosophical Quarterly*, 6, 4: 330-334.

Livadas, Stathis, 2019, "The Plausible Impact of Phenomenology on Gödel's Thoughts, in *Theoria*, 85, 2: 145-170.

—, 2020, "Is there an Ontology of Infinity?", in *Foundations of Science*, 25: 519-540.

—, 2022, "The enigma of 'being there'. Choosing between ontology and epistemology", in *Axiomathes*, 32: 1129–1149.

Pradelle, Dominique, 2021, "On Husserl's Concept of the Pre-predicative: Genealogy of Logic and Regressive Method", in: Engelland, Chad, ed. *Language and Phenomenology*, Routledge, NY: 56-73.

Van Dalen, Dirk, Van Atten, Mark, Tieszen, Richard, 2002, "Brower and Weyl: The Phenomenology and Mathematics of the Intuitive Continuum", in *Philosophophia Mathematica*, 10, 3: 203-226.

Vandevelde, Pol, 2021, "The Scaffolding Role of a Natural Language in the Formation of Thought", in: Engelland, Chad, ed. *Language and Phenomenology*, Routledge, New York: 194-211.

Von Neumann John, 1955, *Mathematical Foundations of Quantum Mechanics*, Princeton University Press, Princeton.

Scientific realism and antirealism in geography

Timothy Tambassi

Abstract: The relationship between (philosophical) scientific (anti)realism and geography is still largely in need of being explored. On one side, the debate on scientific (anti)realism in philosophy of science has led to discussions in and on many scientific disciplines, the list of which rarely includes geographical sciences. On the other side, the geographical debate has outlined its own version of scientific (anti)realism, paying little attention to the literature in philosophy of science. This paper focuses on the geographical literature, with the aim of: 1) showing whether and how the geographical debate is committed to one of the main topics of philosophical scientific (anti)realism, that is: the existence of unobservable theoretical entities; 2) examining the reason(s) why philosophical scientific (anti)realism has been theoretically neglected by geographers. Sect. 2 provides the philosophical framework of our investigation, a framework that, in Sects. 3-6, is used to examine prominent examples from the geographical debate that are explicitly related to ontological analysis. Sect. 7 shows four different reasons why philosophical scientific (anti)realism remains little discussed in geography. Sect. 8, finally, provides some guidelines to enhance communication between geography and philosophy of science on the topic of scientific (anti)realism.

Keywords: philosophy of geography, philosophy of science, scientific antirealism, scientific realism, unobservable entities.

1. *Introduction*

The relationship between (philosophical) scientific (anti)realism and geography is still largely in need of being explored. On one side, the debate on scientific (anti)realism in philosophy of science has led to discussions in and on many scientific disciplines, the list of which rarely includes geographical sciences (see Okasha 2002; Agazzi 2017; Chakravartty 2017; Beebe *et al.* 2020). On the other side, the geographical debate has outlined its own version of scientific (anti)realism, paying little attention to the literature in philosophy of science (Mäki *et al.* 2004). This paper focuses on the geographical literature, with the aim of:

A1. showing whether and how the geographical debate is committed to one of the main topics of philosophical scientific (anti)realism, that is:

ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i2.394
Submitted: 15 November 2021
Accepted: 25 October 2023

the existence of unobservable theoretical entities (Sects. 3-6);

A2. examining the reason(s) why philosophical scientific (anti)realism has been theoretically neglected by geographers (Sects. 7-8).

Concerning A1, Sect. 2 provides the philosophical framework of our investigation. The framework follows the thesis of Corti (2020), holding that the dichotomy between scientific realism [SR] and antirealism [SaR] is independent from the one between metaphysical realism [MR] and antirealism [MaR]. The choice of focusing on Corti's thesis is not random:

1. firstly, it helps to clarify some of the main (philosophical) assumptions behind the dichotomies (see in particular [3] and [4] in Sect. 2), which sometimes are used interchangeably by geographers;
2. secondly, it builds the distinction between SR and SaR on the existence of (un)observable theoretical entities – an existence that this paper aims to discuss within the geographical literature.[1]

In Sects. 3-6, the framework is used to examine prominent examples from the geographical debate that are explicitly related to ontological analysis. More precisely, Sects. 3-4 consider how the nature of geographical entities and the ontological joints of geographical investigation have been discussed. Sect. 5 concerns geographical theories, within which the locution "scientific realism" is mainly associated to Roy Bhaskar's tri-partition of the ontological domains stratifying the world. Sect. 6 discusses the possibility of different SRs and SaRs that are functional to accommodate the peculiarity of the various geographical sub-branches. All those debates are presented by means of numbered lists aimed to reconstruct and isolate the main positions, assumptions, and disciplinary contexts, and to clarify the commitments of geography to (philosophical) SR and SaR. As regards A2, Sect. 7 shows four different reasons why SR and SaR remain little discussed in geography. Sect. 8, finally, provides some guidelines to enhance communication between geography and philosophy of science on SR and SaR. The purpose is thus twofold: reconstructive and speculative. As for reconstruction, this paper offers the first introduction and systematization of philosophical SR and SaR in geography. As for speculation, we think that discussing whether the geographical debate is committed to the existence of unobservable theoretical entities might help geographers to specify the kind(s) of entities they focus on and to clarify some of the theoretical assumptions of geography as a discipline. More generally, the idea is that, since geographers conduct geographical investigations under the guidance of

[1] This does not mean that T2 represents the only way to distinguish SR from SaR. Alai 2017, 2020, for example, claims that the current debate in philosophy of science on SR and SaR is much more focused on the notions of knowledge and justification than on the question of existence of (un) observable theoretical entities.

some theoretical assumptions, for the sake of methodological accuracy, such assumptions should be subject to critical analysis rather than remaining implicit and unexamined.

2. *Between scientific (anti)realism and metaphysical (anti)realism*

In philosophy of science, the question of the existence of unobservable theoretical entities – that is, entities posited by our best scientific theories *and* that human beings cannot observe directly[2] – splits the debate into two main, heterogeneous positions, which do not exclude the chance of views at the boundary between and/or external to them (see Chakravartty 2017; Corti 2020).

[1] SR, in general, claims that (at least some) unobservable entities exist in the same sense in which observable entities such as table, chairs, and so forth do.

[2] SaR, which traditionally includes some forms of empiricism and instrumentalism, does not make any commitment to the existence of unobservable entities.

This means: for sciences that, like paleontology, deal exclusively with observable entities, there is no disagreement between SR and SaR; for sciences that, such as physics or chemistry, make claims about unobservable entities, SR and SaR disagree on the existence of such entities. The disagreement also extends to the general aim of science. While SR argues that science aims to truly describe the world, SaR maintains that providing a true description applies only to the observable part of the world[3] (Okasha 2002).

Now, according to Corti (2020), the dichotomy between SR and SaR should not be, but sometimes is, confused with the one between metaphysical realism [MR] and antirealism [MaR].[4]

[3] MR, in general, claims that:

- [3.1] (a) there exists a mind-independent world, (b) a word that ultimately contains different (kinds of) entities (see Khlentzos 2021);
- [3.2] such a world has a mind-independent structure;
- [3.3] we can know/have access, at least partially, to [3.1(a)], or [3.2], or both.

[4] MaR usually endorses the negation of:

[2] On the distinction between observable and unobservable entities, see Muller 2005; Dicken and Lipton 2006; Turner 2007.

[3] Alternatively, by following van Fraassen 1980, SaR can be taken to hold that science ought to give empirically adequate (in opposition as true, or approximately true) descriptions of the world.

[4] The distinction between MR and MaR does not exclude the chance of positions that consider such a debate as meaningless, unsubstantial and/or unsettled; see for example McDowell 1994; Rosen 1994; Khlentzos 2021.

[4.1] [3.1(a)] and/or [3.2], by considering the world (as well as the entities it contains) as mind-dependent, or dependent on natural languages, human beings' epistemic status, and so forth;

[4.2] (at least) one epistemic claim of [3.3].[5]

In commenting such a metaphysical distinction, Corti (2020: 2) remarks that:

[5] as MR and MaR are umbrella terms covering a wide range of views, they can be divided into different sorts depending on which claims, among [3.1(a)], [3.2], and [3.3] are accepted or rejected. Meaning, each metaphysical (anti) realist should also specify which claims are part of their thesis. Moreover, we cannot fail to emphasize that, because these views may regard different kinds of entities (see [3.1(b)]), it should not be surprising to find out that (anti)realist positions might be independent of each other;

[6] (the independence shown by [5] does not entail that different) (anti)realisms *can(not)* share some connections with other forms of (anti)realism. And this is the case of MR and MaR and of SR and SaR. In other words, it is possible, though not necessary, to hold any combination of scientific and metaphysical realism and antirealism, as Fig. 1 displays.

	MR	**MaR**
SR	[1] + [3]	[2] + [3]
SaR	[2] + [3]	[2] + [4]

Fig. 1. Connections among scientific and metaphysical realism and antirealism.

However, further clarifications are needed to address the question of SR and SaR in the geographical debate, a question that constitutes the main topic of this paper.

First, following Okasha (2002*:* 58-59) and Corti (2020: 3-6), we maintain that the dichotomies between SR and SaR and between MR and MaR should be conceived as logically independent. Claiming the opposite would mean, for example, to exclude the possibility of being scientific realist about unobservable entities without any commitment to a mind-independent external world: a possibility which seems difficult to reject.[6]

Second, following Corti (2020), assuming SR and SaR come in many versions,

[5] For a deeper investigation on the varieties of MR and MaR, see Corti 2020. For an alternative way to present such a dichotomy, see Chalmers 2009; Khlentzos 2021.

[6] For rebuttals to this logical independence, see Psillos 2005; Chakravartty 2017; Massimi 2018; Ladyman 2019, who generally consider SR committed to MR.

[1] and [2] intend to represent only SR's and SaR's minimal assumptions on the existence of observable/unobservable entities.[7] Extending SR and SaR to other assumptions coming from both philosophical and geographical debates might still be possible, although the task is beyond the scope of this paper.

Third, on the basis of [5] and [6], it is not excluded that SR and SaR can have connections with (anti)realist positions other than MR and MaR – and *vice versa*. Fig. 1 should be thus interpreted as insuring their differences as well as their ways of interacting.

3. *Scientific (anti)realism and the philosophical debate on geographical entities*

Before analyzing how (philosophical) SR and MaR have been conceived by geographers, let us spend a few words on the philosophical debate on geographical entities (see Montuschi 2003; Smith 2019; Tambassi 2021), a debate that aims to clarify the nature of entities geographers deal with, and that has so far shown no explicit references to SR and SaR.

In such a debate, the taxonomy of Casati, Smith and Varzi (1998*:* 78-79) represents the only attempt to systematize the different positions at stake. According to the taxonomy, geographical entities are divided into two different sorts, corresponding to the (traditional) dichotomy between physical and human geography. On one side, there are entities such as mountains, rivers, and deserts, whereas, on the other side, there are socio-economic units like nations, cities, and real-estate subdivisions. Starting from this dichotomy, the authors identify three main positions on the existence of geographical entities.

[7] Strong methodological individualism holds that there are no units on the geographic scale, but only people and the tables and chairs they interact with on the mesoscopic level.

[8] Weak methodological individualism claims that, if geographic units exist, they depend or are supervenient upon individuals.

[9] Geographic realism maintains that socio-economic units and other geographic entities have the same ontological standing as the individuals that they appear to be related to.

Establishing whether the distinction among [7-9] has specifically to do with SR or SaR is not that simple. Indeed, from a metaphysical perspective, we can easily argue that conceiving socio-economic units as existing over and above the individuals means that geographical realism assumes MR claiming, in this context, the mind-independence of the geographical reality. Conversely, weak

[7] For an alternative way to present the dichotomy, see Alai 2017, 2020; Massimi 2018.

methodological individualism could represent a position within MaR: in fact, weak methodological individualism does not exclude that (at least part of) geographic reality can be (mind-)dependent upon individuals. (Thomasson 2019: 173), however, would disagree with that: if it is true that MR claims that there are some existing entities which are mind-independent (see [3.1(b)]), there is no reason to think that MR cannot accept that, in addition to those entities, there are also mind-dependent entities in the social world studied, for example, by human geography. Accordingly, weak methodological individualism could be a sort of MR too.) But the distinction among [7-9] makes no explicit reference to unobservable entities. The only thing that we might infer from [7-9] is that, if according to [7] there are only people and the tables and chairs they interact with on the mesoscopic level, then strong methodological individualism does not seem to make any claim about unobservable entities. Thus, strong methodological individualism is not committed to SR. Anything else, from the inclusion of strong methodological individualism within SaR to the inclusion of both geographical realism and weak methodological individualism among SR or SaR, would be, on the basis of [7-9], indeterminate.

4. *Scientific (anti)realism in the geographical debate on ontology*

Sect. 3 has shown that the lack of explicit references to "scientific (anti) realism" in the philosophical debate on geographical entities makes it hard to establish whether the various positions at stake are committed to SR and SaR. The same can be said for the geographical debate on ontology (see Vallega (1995); Berque (2000); Raffestin (2012); Boria (2013)), within which the taxonomy of Tanca (2018) helps to clarify the different views. Such views are categorized according to the "joints that characterize the geographical investigation", namely *things*, *representations*, and *practices*. Those joints, Tanca holds, are independent of each other and correspond to three different lists of ontological claims. The set of all claims in each list outlines one of the different and mutually exclusive ways through which geographers investigate, approach, and interpret the geographical reality.

The ontological claims of the first joint, *things*, are reconstructed as follows.

[10] Geographical reality (and its structure) is mind-independent.

[11] Our knowledge of the geographic reality corresponds to the reality itself.

[12] Sight is the primary and, according to some authors, the *only* means of access to the geographical reality.

[13] Whenever sight alone is not enough, maps and other visual geographical tools can help us in knowing new entities on/of the geographical reality.

As regards the second joint, *representations*, Tanca seems to presume the

following claims.

[14] Geographical reality (and its structure) is mind-dependent, i.e., dependent on our cognitive schemas (that can differ from one another).

[15] Our knowledge of the geographical reality is mediated by language and representation, which, in turn, reflect the social and cultural context within which they are used and have been created.

[16] Languages and representations do not represent mimetically the geographical reality; they shape and create references within and for such a reality.

Finally, the third joint, *practices*, makes the following claims.

[17] Subjects and (geographical) reality affect each other and are (contextually) inseparable.

[18] The geographical reality has a dynamic and processual character that can be explained only be means of the integration of *things*, *representations*, and practices – conceived (the latter) as performances, thought-in-action, and action-in-context. (Practices do not produce entities but constitute subjects' sense of the real).

[19] Subject's knowledge of geographical reality is not exhausted by means of sight, language, and representation, but can be enriched by non-cognitive, expressive, and emotional components of subjects' experience.

Now, if metaphysically speaking, *things* assume MR (see [10] and [11]), *representations* accept MaR (see [14] and [15]), and *practices* seem to deny neither MR nor MaR (see [17]), as regards the dichotomy between SR and SaR, [12] and [13] make the location of the joint *things* difficult. Indeed, on one side, [12] would allow

[20] the inclusion of *things* among SaR because, if sight is the *only* means of access to the geographical reality, then there are no geographical entities that sight cannot see, and therefore *things* do not assume the existence of unobservable entities;

[21] to consider *things* as not committed to SR and SaR: maintaining sight as the only means of access to the geographical reality excludes, in principle, the existence of unobservable entities, and therefore the dichotomy between SR and SaR is not applicable to *things*.

On the other side, [13] seems to enrich the geographical reality with entities that are on maps but cannot be seen with our eyes. Consequently, *things* might be regarded

[22] as a sort of SR, to the extent that [13] does not exclude, at least in principle, unobservable entities: being (observable) on a map, for example, does not mean being observable *per se* (see, for instance, Sandy Island);

[23] as not committed to SR and SaR, insofar as [13] does not specify whether unobservable entities exist in the same sense in which observable entities do.

But the same argument can also be extended to *representations* and *practices*, by replacing, in [22] and [23], [13] respectively with [16] and [19].

5. *"Scientific (anti)realism" in the geographical debate*

The lack of references to the locution "scientific (anti)realism" in the debates of Sects. 3-4 does not imply, however, that the locution has never been mentioned in the whole geographical investigation. However, as Mäki and Oinas (2004) remark, such a locution appears rarely in geography, and when it does, little attention is given to the literature on SR and SaR in philosophy of science. Moreover, unlike the philosophical debate, within which the dichotomy between SR and SaR questions the existence of unobservable entities (see Sect. 2), the geographical investigation often connects "scientific (anti) realism" to:

[24] the notion of observation, without references to the debate on unobservable entities (Yeung 1997; Brown 2004);

[25] the question of causality *and* the notion of observation (Lawson and Staeheli 1990);

[26] the existence of mind-independent and/or mind-dependent reality in physical and human geography, both from a metaphysical and epistemological perspective (see [3] and [4]) (Harrison and Livingstone 1979; Mäki and Oinas 2004).

Few exceptions occur which generally refer to the question of unobservable entities within Bhaskar's scientific realism (1975a, 1975b, 1979, 2009),[8] with emphasis on social research (Sarre 1987), international relations theory (Wendt 1987), different kinds of realism in geography (Rose 1990), social construction of the notion of nature (Proctor 1998), physical geography (Tucker 2009), and middle power scholarships (Jeong 2019). About how the question of unobservable entities fits into Bhaskar's scientific realism, Jeong (2019: 248-249) makes the point clear, maintaining that, for Bhaskar's scientific realism, the world is stratified in three different ontological domains:

[27] the real, which includes the things that exist, and their structure and power, known as mechanisms;

[28] the actual that comprehends observable or unobservable events generated by those mechanisms when activated;

[8] Bhaskar's scientific realism has been proposed in two different versions: transcendental realism for natural sciences and critical realism for social sciences (Yeung 1997). Other authors of references for scientific (anti)realism in geography are Keat and Urry 1975 and Sayer 1982a, 1984, 1985a, 1985b, 1992, 2000, who, just as Bhaskar, are rarely mentioned in the debate on SR and SaR in philosophy of science.

[29] the empirical, that is, events we experience/observe directly or indirectly.

On this basis, Jeong (2019) further specifies that:

the fundamental assumption [of this kind of scientific realism] *is that there is a world independent of human thought*, and to understand such a world requires two different dimensions of science: the "transitive" and "intransitive". The intransitive dimension holds the relatively unchanging things of the world, "the object of science [...] in the sense of the things we study – physical processes or social phenomena". The transitive dimension is formed through theories and methods concerning the objects of study in the intransitive dimension. So, while different theories and methods seek to explain the objects in the intransitive dimension, those very objects of study remain the same. Theories and methods may change or be replaced over time, but that does not necessarily mean the objects also change. [...] *The investigation of that intransitive social world can reveal* [...] *features unobservable in the domain of the empirical*» (Jeong 2019: 249, emphasis added).

This means, according to [28] and to the quotation, Bhaskar's scientific realism assumes [1] (but also [3]), to the extent that unobservable events and features are explicitly not rejected.[9] About the assumptions, Montuschi adds more details:

according to [Bhaskar's] model, *scientific objects are* ontologically "intransitive" (existing independently of our knowledge/methods of inquiry) and *unobservable* (*conceived in terms of generative mechanisms or structures, of which empirical, observable phenomena are only a manifestation*). The difference between natural and social objects [...] consists of the type of independence they have from knowledge/inquiry: it is total independence, in the case of the former; partial, in the case of the latter. [...] *Social objects, unlike natural ones, do not exist independently of the activities they govern (and also they cannot be identified independently of them empirically).* [...] Social objects – unlike natural ones – do not exist independently of the agents' conceptions of what they are doing in their activities. [...] This also means that social objects are 'conceptualized in the experience of the agents concerned' and since people's conceptualizations have a history, these objects are not immutable (marriage, like any other institution, can change over time). Finally, and more generally, it has to be acknowledged that the social sciences, unlike the natural sciences, are part of their own field of inquiry, in the sense that they are 'internal' with respect to their subject matter. This makes social scientific categorizations self-referential, and the referents of social scientific inquiry

[9] See Sarre 1987; Rose 1990. Moreover, according to Bhaskar 1975a, [29] should be considered as committed to unobservable entities. In Bhaskar 1979, it is also pointed out that science seeks causal laws to explain observed events, and that these causal laws deal with tendencies in objects, some of which may be unobservable.

themselves dependent on the processes which produce the knowledge of those very referents. Nonetheless, *partial independence is only taken to demonstrate that the objects of social science are of a specific nature* (i.e. social nature), *not that they do not constitute a category of scientific objects – and even less, that they cannot be treated scientifically* (Montuschi 2003: 14-15, emphasis added).

In a nutshell, both natural and social entities, which Montuschi links to physical and human geography respectively, are unobservable. And if it is true that social entities depend on and cannot be *empirically* identified independently to the activities they govern, it should also be emphasized that the "empirical" in question refers to [29], that is, one of the three ontological domains stratifying the world. According to Rose (1990: 166, 169), such domains exhibit, for Bhaskar's scientific realism, the following relationship, [27] > [28] > [29],[10] and altogether postulate the existence of entities that are unobservable in character.

While, in the geographical debate, the importance of Bhaskar's scientific realism has been highlighted by authors such as Gregory (1978, 1982), Sayer (1982b, 1987) and Cooke (1987), there has been also criticism, especially as regards the postulation of unobservable entities. Sack (1982), for example, asks whether those entities really exist or are product of our own theorizing. Allen (1987), instead, affirms that unobservable entities are just conjectured and if they existed, they would have accounted for certain types of events. But since there is no guarantee that such entities do exist, Bhaskar's scientific realism remains wholly hypothetical and dependent on the a priori transcendental argument (Rose 1990: 169).

6. *Scientific (anti)realisms and geographical sub-branches*

Sala (2009) splits the whole geographical domain in three different areas of research, namely: human, physical, and technical geography. Going back in time, Pattinson (1963) distinguishes four historical traditions within the geographical investigation: spatial, area studies, man-land, and earth science. More recently, Agnew and Livingston (2011) and Johnston and Sidaway (2016) map the disciplinary space of geography as a set of movements, flows, and channels, by identifying over a dozen geographical schools. None of these classifications excludes the possibility of including, among its ramifications, the multiplicity of sub-branches characterizing the geographical investigation – i.e., economic, social, tourism, coastal, bio-, hydro, transportation geography, and so forth.

[10] Where [27] is meant to include [28] that, in turn, includes [29].

The heterogeneity of the geographical investigation, including the fact that geographical sub-branches have particular lists of entities (see Sect. 3 and Tambassi 2021) and that (anti)realist positions may be independent of each other (see [5]), pushes Mäki and Oinas (2004: 1772) towards a general skepticism about a global (anti)realism capable of accommodating all geographical sub-branches in a suitable and profitable way (see also Mäki 1996). Instead, they propose a series of local (anti)realisms, each tailored to grasp the view of a certain discipline (such as realism about geography, realism about biochemistry, realism about archaeology, and so forth) or even smaller units such as specific research fields and theories, like cultural geographies, coastal geography, and so on. As for SR and SaR, Mäki and Oinas maintain that any local (scientific anti)realisms should meet two main constraints:

[30] the specific contents of any (anti)realism, that is, for SR and SaR, the question of the existence of unobservable theoretical entities (see [1] and [2]);

[31] the peculiar features of the local discipline.

It is on the basis of such constraints that Rhoads and Thorn (1994) have focused on the *potential* contribution of (philosophical) SR and SaR, as they are conceived in Sect. 2, to geomorphology (but also to others sub-fields of the physical geography). On one side, they argue that the challenge is to show how many theoretical constructs embodied in geomorphology, including references to unobservables, have been preserved in contemporary geomorphic theories (see [30]). On the other side, by following [31], they contend that

> there is no reason to presuppose that a philosophical framework for geomorphology will be merely a restatement of the philosophy of another discipline. Because geomorphology is concerned with distinctive types of natural systems that include synergistic physical and biological elements and employs characteristic investigative methods, it cannot be reduced to the underpinning disciplines (Rhoads and Thorn 1994: 98).

7. *Four reasons for a theoretical marginality*

The theses presented so far are isolated cases in the geographical debate, within which the question of the existence of unobservable theoretical entities generally remains marginal. The reasons for this, geographers suggest, are essentially four: the first two, [R1] and [R2], concern the reception of philosophical scientific (anti)realism(s) in geography, the last two, [R3] and [R4], explicitly question the issue of unobservable entities.

R1 can be tracked in the words of the majority of geographers above and refers to the relationship between geography and philosophical scientific (anti)

realism. While Brown (2004: 369) affirms that SR (and SaR) should not be confused with any particular philosophy, Yeung (1997: 51) emphasizes that SR (and SaR) is a philosophy intrinsically, and Sack (1982: 504) adds that geography should not make its questions (and methods) adhere to philosophy, but rather use philosophy to help focus on geographical questions. Accordingly, the question of the existence of unobservable theoretical entities rarely appears in the geographical debate because it is just a philosophical question and not a geographical one.

R2 specifically refers to the thesis of Yeung (1997), according to which the marginality of such a question stems from the multiplicity of realisms populating the geographical investigation – a multiplicity that has confined the question of unobservable entities into the background, lost among the various questions emerging from different realisms (in geography).

> [T]he crux of most recent debates in [...] geography rests upon a misreading of different moments of [scientific] realism. There seems a lack of proper understanding of [scientific] realism in its own terms. Critics of realism and realist research in human geography rely largely upon cursory readings of different versions of 'realism' presented in the geographic literature. It is not surprising that many of them are confused between treating [scientific] realism as a philosophy, an epistemology, a method, a dogma or just another '-ism' (Yeung 1997: 54).

R3, instead, connects the topic of unobservable entities to the level of granularity geographers refer to. Sect. 2 has emphasized that some sciences make claims about unobservable entities, whereas other sciences do not. According to Smith and Klagges (2009), such claims may depend on the levels of granularity of different scientific investigations. Since, by following Egenhofer and Mark (1995), the level of granularity of geography coincides with the mesoscopic stratum[11] of spatial reality and includes entities such as «Vienna, with its streets, buildings, parks, and people», «Europe with mountains, lakes and rivers, transportation systems, political subdivisions, cultural variations, and so on», there is no room for unobservable entities in the geographical investigation. In other words, the question of the existence unobservable theoretical entities remains marginal because, from a geographical point of view, SR and SaR cannot disagree on their existence.

R4, in contrast to R3, does not deny the chance of references to unobservable entities within the geographical domain. Sect. 2 has highlighted that un-

[11] Geographically speaking, Egenhofer and Mark maintain that the mesoscopic stratum represents the space where we move. Such a space is distinct from the small-scale space, populated by objects and events smaller than those that can easily be seen by the naked eye.

observable theoretical entities are entities that human beings cannot observe directly. In this regard, Tucker (2009) spots that those entities may refer to a wide variety of phenomena, which include objects unobservable for their size (atoms), or that are somehow hidden (the core of planet earth), or too distant from us (black holes). But Tucker further suggests that "unobservable entities" in geography may also refer to events of the past, which are unobservable because they are distant in time. And this (last) perspective seems to be adopted by Inkpen and Wilson (2013) to describe physical geography and earth sciences (also) in terms of historical disciplines that make hypotheses about unobservable (past) events, and by Tanca (2018) for (further) distinguishing the joint *things* from the joint *representation* (see Sect. 4): the former conceives geographical entities in terms of current existence, whereas the latter does not disregard their past. But the list of unobservable theoretical entities in geography is also enriched by Lawson and Staeheli (1990: 13), who include "the unseen social structures", which influence, and are influenced by, the actions of individuals. On this basis, the question of the unobservable entities would be thus marginal in geography because the references for those entities are fuzzy and may vary from context to context.

8. *Final remarks*

According to Corti (2020: 3), one of the main issues of SR and SaR is that both are umbrella terms: such an issue is so widespread that introducing oneself as a scientific (anti)realist is too vague, if the (kind of) (anti)realism at stake is no further specified. In Sect. 2, this vagueness has been reduced by means of the identification, in [1] and [2], of the minimal claims from which all the variants of SR and SaR start building their views (Corti 2020: 6), without denying that such variants can be enhanced by other claims. On the basis of such claims, this paper has analyzed the reception of (philosophical) SR and SaR within the geographical investigation. Sects. 3-4 have shown that, although SR and SaR are not explicitly mentioned, the different positions on the existence of geographical entities and on the joints characterizing the geo-ontological investigation might be somehow committed to [1] and [2]. Sect. 5 has underlined that, when the locution "scientific (anti)realism" appears in geography, it is mainly associated to Bhaskar's scientific realism, which absorbs the question of unobservable entities posited by [1] and [2] within the three different ontological domains (see [27-29]) stratifying the world. Sect. 6, in contrast, has placed [1] and [2] at the center of the geographical debate, and has suggested that SR and SaR may vary depending on the geographical sub-branches they refer to. Sect. 7 has, finally, emphasized four different reasons why the question

of unobservable entities remains marginal in the geographical debate, namely: the difficult relationship between geography and realism (R1), the multiple and overlapping (anti)realisms populating the geographical investigation (R2), the level of granularity geographers refer to (R3), and the ambiguous references for unobservable entities in geography (R4).

All those considerations could still provide some guidelines to enhance communication between geography and philosophy of science about the question of unobservable entities. The distinction between [1] and [2], for example, would allow us to meet the concern raised in R2: clarifying the claims of SR and SaR may, in fact, reduce the lack of proper understanding of (philosophical) scientific (anti)realism in geography, in its own terms. But the same distinction could also shed light to R1, by considering SR (and SaR) as exclusively philosophical and not geographical. Geography would not, however, be excluded from the debate of unobservable entities: by reconsidering [30] and [31], we could maintain that, while the question of [30] is philosophical, the domain which [30] refers to – that is, [31] – is geographical and should meet the peculiar features of the different geographical investigations. And such features cannot but consider, according to Rhoads and Thorn (1994) and to R4, that the references for unobservable entities in geography may vary depending on the geographical sub-branch (or the theory) we analyze. This means, different geographical sub-branches can refer to different kinds of unobservable entities, not excluding that some of those sub-branches might also leave no room for unobservable entities (see R3). Our proposal is thus close to the thesis of Mäki and Oinas (2004), who suggest a multiplicity of local SRs and SaRs accommodating the needs of different disciplines or even of their sub-branches and particular theories. To [30] and [31], which already apply Mäki's and Oinas' proposal to the specificity of SR and SaR, we could add a further constraint aimed at explicitly remarking that:

[32] references for unobservable entities can vary depending on the geographical sub-branch or theory we deal with, by exhibiting specific peculiarity.

Timothy Tambassi
Ca' Foscari University of Venice
timothy.tambassi@gmail.com

References

Agazzi, Evandro, 2017, ed., *Varieties of Scientific Realism. Objectivity and Truth in Science*, Springer, Cham.

Agnew, John A., *et al.*, 2011, eds., *The SAGE Handbook of Geographical Knowledge*, SAGE Publications, London.

Alai, Mario, 2017, "The Debates on Scientific Realism Today: Knowledge and Objectivity in Science", in Agazzi [2017]: 19-47.

—, 2020, "Scientific Realism, Metaphysical Antirealism and the No Miracle Arguments" in *Foundation of Science* 28: 377-400.

Allen, John, 1987, "Realism as Method" in *Antipode* 19, 2: 231-239.

Beebe, James R., *et al.*, 2020, "Scientific Realism in the Wild: An Empirical Study of Seven Sciences and History and Philosophy of Science" in *Philosophy of Science* 87: 336-364.

Berque, Augustin, 2000, *Mediance: de milieux en paysages*, Belin, Paris.

Bhaskar, Roy, 1975a, "Forms of Realism" in *Philosophica* 15, 1: 99-127.

—, 1975b, *A Realist Theory of Science*, Leeds Books, Leeds.

—, 1979, *The Possibility of Naturalism*, Harvester, Hassocks.

—, 2009, *Scientific Realism and Human Emancipation*, Routledge, London-New York.

Boria, Edoardo, 2013, "Genealogie intellettuali e discontinuità nazionali nella storia della cartografia" in *Bollettino della Società Geografica italiana* 6, 3: 443-460.

Brown, James D., 2004, "Knowledge, Uncertainty and Physical Geography: Towards the Development of Methodologies for Questioning Belief" in *Transactions of the Institute of British Geographers* 29, 3: 367-381.

Casati, Roberto, *et al.*, 1998, "Ontological tools for geographic representation" in Guarino, ed., *Formal ontology in information systems*, IOS Press, Amsterdam: 77-85.

Chakravartty, Anjan, 2017, "Scientific realism" in Zalta, ed., *The Stanford encyclopedia of philosophy*, https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/.

Chalmers, David J., 2009, "Ontological anti-realism" in Chalmers *et al.*, eds., *Metametaphysics: New essays on the foundations of ontology*, Oxford University Press, Oxford.

Cooke, Philip, 1987, "Individuals, Localities and Postmodernism" in *Society and Space* 5, 4: 408-412.

Corti, Alberto, 2020, "Scientific Realism Without Reality? What Happens When Metaphysics is Left Out" in *Foundation of Science* 28: 255-475.

Dicken, Paul *et al.*, 2006, "What can Bas Believe? Musgrave and van Fraassen on Observability" in *Analysis* 66, 291: 226-233.

Egenhofer, Max J. *et al.*, 1995, "Naive geography" in Frank *et al.*, eds., *Spatial information theory: a theoretical basis for GIS* in *Proceedings of the second international conference*, Springer, Berlin-Heidelberg: 1-15.

Gregory, Derek, 1978, *Ideology, Science and Human Geography*, Hutchinson & Co. Press, London.

—, 1982, "A Realist Construction of the Social" in *Transactions: Institute of British Geographers* 7: 254-256.

Harrison, Richard *et al.*, 1979, "There and Back Again: Towards a Critique of Idealist Human Geography Source" in *Area* 11, 1: 75-79.

Inkpen, Robert *et al.*, 2013, *Science, philosophy and physical geography*, Routledge, London.

Jeong, Monica S., 2019, "Critical realism: A better way to think about middle powers" in *International Journal: Canada's Journal of Global Policy Analysis* 74, 2: 240-257.

Johnston, Ron *et al.*, 2016, eds., *Geography & Geographers. Anglo-American Human Geography since 1945*, Routledge, London-New York.

Keat, Russell, 1975, *Social Theory as Science*, Routledge - Kegan Paul, London.

Khlentzos, Drew, 2021, "Challenges to metaphysical realism" in Zalta, ed., *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2021/entries/realism-sem-challenge/.

Ladyman, James, 2019, "What is the Quantum Face of Realism" in Lombardi *et al.*, eds., *Quantum Worlds. Perspectives on the Ontology of Quantum Mechanics*, Cambridge University Press, Cambridge: 121-132.

Lawson, Victoria *et al.*, 1990, "Realism and the practice of geography" in *The Professional Geographer* 42: 13-20.

Mäki, Uskali, 1996, "Scientific realism and some peculiarities of economics" in *Boston Studies in the Philosophy of Science* 169: 425-445.

—, *et al.*, 2004, "The narrow notion of realism in human geography" in *Environment and Planning* A 36: 1755-1776.

Massimi, Michela, 2018, "Four kinds of perspectival truth" in *Philosophy and Phenomenological Research* 96, 2: 342-359.

McDowell, John, 1994, *Mind and World*, Harvard University Press, Cambridge.

Montuschi, Eleonora, 2003, *The Object of Social Science*, Continuum, London-New York.

Muller, Fred, 2005, "The Deep Black Sea: Observability and Modality Afloat" in *British Journal for the Philosophy of Science* 56, 1: 61-99.

Okasha, Samir, 2002, *Philosophy of science: A very short introduction*, Oxford University Press, Oxford.

Pattinson, William D., 1963, "The four traditions of geography" in *J Geogr,* 63 5: 211-216.

Proctor, James D., 1998, "The Social Construction of Nature: Relativist Accusations, Pragmatist and Critical Realist Responses" in *Annals of the Association of American Geographers* 88, 3: 352-376.

Psillos, Stathis, 2005, "Scientific realism" in *Encyclopedia of philosophy*, Gale Macmillan Reference, Farmington Hills.

Raffestin, Claude, 2012, "Space, Territory, and Territoriality. Environment and Planning" in *D: Society and Space* 30, 1: 121-141.

Rhoads, Bruce L., *et al.*, 1994, "Contemporary Philosophical Perspectives on Physical Geography with Emphasis on Geomorphology" in *Geographical Review* 84, 1: 90-101.

Rose, Courtice, 1990, "Toward Pragmatic Realism in Human Geography" in *Cahiers de géographie du Québec* 34, 92: 161-179.

Rosen, Gideon, 1994, "Objectivity and modern idealism: What is the question?" in Michael *et al.* (eds.), *Philosophy in mind*, Kluwer Academic Publishers, Dordrecht: 277-319.

Sack, Robert David, 1982, "Realism and Realistic Geography" in *Transactions: Institute of British Geographers* 7: 504-509.

Sala, Maria, 2009, "Geography" in Sala (ed.), *Geography. Encyclopedia of life support systems*, EOLSS Publisher, Oxford: 1-56.

Sarre, Philip, 1987, "Realism in practice" in *Area* 19, 1: 3-10.

Sayer, Andrew, 1982a, "Explanation in economic geography" in *Progress in Human Geography* 6, 1: 68-88.

—, 1982b, "Misconceptions of Space in Social Thought" in *Transactions: Institute of British Geographers* 7: 494-503.

—, 1984, *Method in Social Science: A Realist Approach*, Routledge, London.

—, 1985a, "Realism and geography" in Johnston (ed.), *The Future of Geography*, Methuen, London: 159-173.

—, 1985b, "The difference that space makes" in Urry (eds.), *Social Relations and Spatial Structures*, Macmillan, London: 49-66.

—, 1987, "Hard Work and Its Alternatives" in *Society and Space* 5, 4: 395-399.

—, 1992, *Method in Social Science: A Realist Approach*, Routledge, London.

—, 2000, *Realism and Social Science*, Sage, London.

Smith, Barry, 2019, "Drawing Boundaries" in Tambassi (ed.), *The Philosophy of GIS*, Springer, Cham: 137-158.

—, *et al.*, 2008, "Bioinformatics and philosophy" in Munn *et al.* (eds.), *Applied ontology. An introduction*, Ontos-Verlag, Berlin: 21-37.

Tambassi, Timothy, 2021, *The Philosophy of Geo-ontologies. Applied Ontology of Geography*, Springer, Cham.

Tanca, Marcello, 2018, "Geografia e filosofia: istruzioni per l'uso" in *Semestrale di Studi e Ricerche di Geografia* 30, 2:13-27.

Thomasson, Amie, 2019, "Geographic Objects and the Science of Geography" in Tambassi (ed.), *The Philosophy of GIS*, Springer, Cham: 159-176.

Tucker, Aviezer, 2009, "The Philosophy of Natural History and Historiography" in *Journal of the Philosophy of History* 3: 385-394.

Turner, Derek Donald, 2007, *Making Prehistory: Historical Science and the Scientific Realism Debate*, Cambridge University Press, Cambridge.

Vallega, Adalberto, 1995, *La regione, sistema territoriale sostenibile: compendio di geografia regionale sistematica*, Mursia, Milano.

van Fraassen, Bas, 1980, *The Scientific Image*, Oxford University Press, Oxford.
Wendt, Alexander, 1987, "The Agent-Structure Problem in International Relations Theory" in *International Organization* 41, 3: 335-370.
Yeung, Henry Wai-chung, 1997, "Critical realism and realist research in human geography: a method or a philosophy in search of a method?" in *Progress in Human Geography* 21, 1: 51-74.

Focus

Introduction: 15 years of discussion on moral enhancement

Sergio Filippo Magni, Elvio Baccarini

1. *For and against moral enhancement*

Improvement in knowledge of the neurobiological bases of behavioural disposition with moral relevance have stimulated ethical reflection on the opportunity to employ biotechnological devices and resources to improve human morality (Clarke *et al.* 2016). Discussion on biotechnological moral enhancement, as a separated issue from that of biotechnological cognitive human enhancement, has started after the publication of two seminal articles in 2008: "The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity" written by Ingmar Persson and Julian Savulescu, and "Moral Enhancement" written by Thomas Douglas.

Since that moment, bioethical debate on this topic has been divided in two separate fields: anti-moral-enhancement views, which consider the attempt of morally enhancing human beings as immoral, and pro-moral-enhancement views, which consider the attempt of morally enhancing human beings as moral. In both sides there are more radical and less radical views.

Radical anti-moral-enhancement views refuse any kind of moral enhancement as well as any kind of cognitive enhancement aimed to improve agent's knowledge and rationality. According to such a bio-conservative view, any attempt to enhance human nature beyond its normal biology is claimed to be perfectionist and dehumanizing. Examples are Michael Sandel (2007) or Leon Kass (2008). On the other hand, moderate anti-moral-enhancement views accept cognitive enhancement as a means to increase human capacity of reasoning and acting but refuse moral enhancement. We are morally justified to enhance human beings by improving their cognitive abilities, but not by intervening on their moral motivations. As John Harris states, quoting Milton's "Paradise Lost", human beings ought to remain "free to fall", free to do immoral things: "there are substantial issues of liberty […] which could conceivably be threatened by any measures that make the freedom to do immoral things impossible, rather than simply making the doing of them

philinq XI, 2-2023, pp. 53-56
ISSN (print) 2281-8618-ETS

wrong and giving us moral, legal, and prudential reasons to refrain" (Harris 2016: 64).

On the contrary, radical pro-moral-enhancement views consider moral enhancement morally obligatory: an action that the agent ought to perform. An example is Persson&Savulescu (2012, 2: "we shall contend that in order for the majority of citizens of liberal democracies to be willing to go along with constramts on their extravagant consumption, their moral motivation must be enhanced so that they pay more heed to the interest of future generations and non-human animals". On the other hand, moderate pro-moral-enhancement views consider moral enhancement as merely morally permitted: an action that the agent is free to perform or not to perform. An example is Douglas (2008: 233): "I will tentatively argue that it would sometimes be morally permissible for people to biomedically mitigate their counter-moral emotions".

2. *Varieties of moral enhancement*

Yet, the picture of pro-moral-enhancement positions arisen the last 15 years is much more complex. Indeed, there are many kinds of moral enhancement, and all these kinds may be accompanied by different moral evaluations.

Firstly, moral enhancement can be direct or indirect. Properly, what is at issue in the debate is not indirect moral enhancement (the common way to slowly improve human moral motivations through education, culture, examples and so on: these traditional enhancers are usually considered morally right), but only direct moral enhancement, the new possibilities of fast improving human moral motivations using biotechnologies (mental drugs, genetic engineering and so on: the new moral enhancers), which is much more controversial.

Secondly, moral enhancement can be negative or positive. Negative, if morally enhancing human beings is directed to eradicate anti-social motivations, connected to emotions like anger, hate, aggressivity etc. (Douglas 2008; 2013 are an example of this kind of moral enhancement). Positive, if morally enhancing human beings is directed to enforce pro-social motivation, connected to attitudes like altruism, sympathy, sense of justice (Persson *et al.* 2008; 2012 are an example of this kind of moral enhancement). Thirdly, moral enhancement can be voluntary or involuntary, according to whether a person wants or does not want to be morally enhanced. But it can also be compulsory, because performed against the person's will.

All these distinctions are enough to clarify why such a matter can be considered a very contentious problem. But the problem seems to be even more contentious when we add other distinctions useful to complete the picture. Moral enhancement can be internal or external, according to whether it is di-

rected to modify agent's moral mental states (attitudes, emotions, motivations and so on), or it is directed to create external conditions that can affect agent's moral decisions (because of technological devices, artificial intelligence and so on). Moreover, moral enhancement can be procedural or substantive, according to whether it regards the mental process of elaborating moral decisions or the very content of these decisions: how individuals arrive at deciding to do the good, or what individuals endorse to be good. Finally, moral enhancement can be specific or general, depending on whether it is intended to be confined to a particular set of people (violent people, psychopaths, murderers and so on) or to be directed toward every people: to all the human beings.

3. *The articles in the Focus*

The elaboration of such a picture and these latter distinctions are helpful to introduce the articles of our focus. Matteo Galletti' article, "Internal and External Moral Enhancements: The Ethical Parity Principle and the Case for a Prioritization", separates the moral evaluation of internal and external enhancement, giving priority to the internal one, and defends this position from the critique made by Neil Levy, who has endorsed the so-called Ethical Parity Principle between internal and external enhancement. In "Creating Capabilities to be Better", Francesca Guma endorses a kind of procedural moral enhancement directed to improve human free-will, intended, in the light of a distinction originally posed by J.L. Austin's, as opportunity and capacity to will otherwise. Both the articles deal with a kind of general moral enhancement, potentially directed toward every human being.

On the other hand, in "Public Reason and Biotechnological Moral Enhancement of Criminal Offenders", Elvio Baccarini defends a kind of specific moral enhancement directed to criminal offenders, based on a Rawlsian method of public reason. Such a method could justify the legitimacy of the proposal to use biotechnology to perform a moral enhancement of people who have committed serious crimes and represent a persistent danger to society. According to Baccarini, such a compulsory moral enhancement against the agent's will could be legitimate, but it must be publicly justified by reasonable agents.

Sergio Filippo Magni
University of Pavia
filippo.magni@unipv.it

Elvio Baccarini
Universiy of Rijeka
ebaccarini@ffri.hr

References

Clarke, Steve *et al*., 2016, eds., *The Ethics of Human Enhancement: Understanding the Debate*, Oxford University Press, Oxford.

Douglas, Thomas, 2008, "Moral Enhancement" in *Journal of Applied Philosophy* 25: 228-245.

Douglas, Thomas, 2013, "Moral Enhancement though Direct Emotional Modulation. A Reply to John Harris" in *Bioethics* 27: 160-168.

Harris, John, 2016, *How to be Good: The Possibility of Moral Enhancement*, Oxford University Press, Oxford.

Kass, Leon, 2008, "Defending Human Dignity" in President's Council on Bioethics, ed., *Human Dignity and Bioethics. Washington DC: US Government Printing Office*: 297-332.

Persson, Ingmar, *et al*., 2008, "The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity" in *Journal of Applied Philosophy* 25: 162-177.

—, *et al*., 2012, *Unfit for the Future: The Need for Moral Enhancement*, Oxford University Press, Oxford.

Sandel, Michael, 2007, *The Case Against Perfection. Ethics in the Age of Genetic Engineering*, Belknap, Harvard.

Internal and external moral enhancements: the ethical parity principle and the case for a prioritization

Matteo Galletti

Abstract: Is there any moral difference between internal moral enhancements, which directly affect the biological nature of human beings, and external moral enhancements, which nudge choices and behavior without changing human biology? If Neil Levy's Ethical Parity Principle is applied, the answer should be no. Recently, John Danaher has argued that the Ethical Parity Principle is invalid and that there are ethical and political reasons for a prioritization of internal over external moral enhancements. Although Danaher's argument presents some interesting insights, it needs to be corrected with finer-grained distinctions of the types of moral enhancements.

Keywords: moral enhancement, behavioral ethics, nudge, procedural moral enhancement, ethical parity principle.

1. *Moral Enhancements: Internal and External*

In the debate on moral enhancement, one of the proposed but little-discussed distinctions is that between internal and external enhancements. Before introducing it, however, I need to clarify what moral enhancement means. Following some suggestions from DeGrazia (2013), I propose this definition: an enhancement is any deliberate intervention that strengthens or reduces existing capacities and dispositions or creates new ones to improve the motivation, decision-making, and behavior of an individual or population in the moral domain.

This definition raises some questions. First, it includes under the label of moral enhancement interventions that do not increase moral traits and capacities but mitigate or eliminate certain tendencies deemed to be pernicious, such as dispositions to violent reaction, implicit biases, or, generically, "counter-moral" emotions (Douglas 2008). Suppose an effect is to improve an individual's moral condition. In that case, we consider it indifferent whether the way by which it is achieved is active (increase or reinforcement) or negative (erasure or reduction).

Second, according to some authors, a moral enhancement qualifies as such because of its effect, regardless of the intentions with which it is implemented. I find this objection reasonable, but the reference to intentionality allows

philinq XI, 2-2023, pp. 57-72
ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i2.492

for a distinction between "enhancement" and "improvement", which I think is relevant to the judgment of this kind of technique. An "improvement" occurs when a given intervention (increase or decrease) on some property of the organism betters its condition so that at instant t0 (prior to the enhancement), the condition of the organism is X, and at instant t1 (after the enhancement) the condition of the organism is Y, where Y is judged better than X. An enhancement aimed at improving could be causally effective in reducing or increasing a certain capacity but not result in an actual improvement in the individual's condition: one might think that an enhancement that endows an individual with the stature of 3 meters would provide the individual with a positional good because he or she will have an advantage in some activities (e.g., sports), but might adversely affect many other areas of his or her life, given the difficult adaptability of his or her stature to the surrounding environment. Alternatively, the enhancer, aware of this outcome, might practice such an intervention precisely with the intention of harming the individual. The distinction between "enhancement" and "improvement" allows finer-grained judgments.

Third, the triad being addressed (judgment, motivation, behavior) includes some rather heterogeneous elements of morality (judgment and motivation belong to psychology, and behavior indicates an observable external expression). Nevertheless, the central distinction here is between the term "dispositions", which includes character traits and moral dispositions such as altruism or empathy, and the concept of "capacities", which refers to second-order reflective capacities, such as moral reasoning, deliberation, and imagination. I will return to this distinction later because it is revealing for understanding the moral status of various types of enhancement.

If we accept this general definition of "moral enhancement", an internal enhancement is a deliberate intervention that either strengthens or reduces existing capacities and dispositions or creates new ones by acting directly on biology, with the aim of improving the motivation, decision-making, and behavior of an individual or population in the moral domain. For example, drug administration and genome editing intervention on the somatic or germline that have this effect can be considered internal enhancements. Thus, internal enhancements involve integrating the biotechnological intervention into the organism's biology. I propose to reserve the name "moral bio-enhancements" (MBE) for these interventions.

External enhancements consist of other means of improving moral traits and abilities, such as external devices, without directly affecting or integrating with the organism's biological constitution; some other external enhancements introduce changes in the context to achieve the enhancing effect. Examples of external enhancements are the use of artificial intelligence devices to make moral

decisions (Borenstein *et al.* 2016; Giubilini *et al.* 2018; Lara 2021) or specific changes in the context of choice that affect sensory perception and favor certain moral judgments (Schnall *et al.* 2008; Wheatley *et al.* 2005; Eskine *et al.* 2011). I propose to call these interventions "moral environment-enhancements" (MEE).

As I understand it, the distinction between MBE and MEE does not perfectly overlap with the distinction between biotechnological enhancements and so-called "traditional" enhancements, such as education, socialization, and the organization of a system of rewards and punishments. Traditional enhancements can, at best, be considered as a species of the MEE genre, including advanced technological interventions such as AI devices or, as we shall see, choice architectures. It is precisely the behavioral sciences that are helping to provide a description of moral agency whose natural limits necessitate the use of innovative measures to influence ex-ante the actions of individuals. In this essay, we will focus on a particular type of these external enhancements, namely so-called "nudges", to induce morally desirable behavior in individuals.

The literature on moral enhancement has only occasionally considered the distinction between external and internal moral enhancements. In this essay, we will argue that any judgment of the moral superiority of MBEs over MEEs, such as moral nudges, cannot be general in nature but must be circumstantial. We will consider Danaher's recent contribution to the debate.

2. *Behavioral Ethics and Moral Nudges*

The numerous empirical research in psychology and behavioral economics have defined a field of studies, which includes rather heterogeneous approaches to the phenomenon of morality, called "behavioral ethics" (BE) (Bazerman *et al.* 2012). BE "addresses people's inability to fully recognize the ethical, moral and legal aspects of their behavior" (Feldman 2018: 2); although BE shares with behavioral analysis the empirically supported belief that biases affect individual choices, it departs from it concerning the general explanation of how these biases work. According to the behavioral sciences, biases are due to the involvement of automatic responses, unmediated by reflexivity and deliberative reasoning, that take the form of post hoc rationalization to justify unethical behavior. Instead, BE provides a more complex picture of moral agency, in which the limitations of various cognitive capacities are due to the tendency to seek self-interest and the inherent need to maintain a coherent and positive self-representation. The situation in which the agent chooses also has a limiting impact on perception, judgment, and choice. The action of these mechanisms occurs mostly unconsciously and may also resort to post-hoc strategies of moral disengagement, thus leading to a hiatus between full personal awareness of what the

agent is doing and the actual intention to do harm. In general, ethical biases inhibit individuals' ability to recognize the moral quality of their actions (35-36).

Innovative tools for intervening in human behavior (88-98) include nudges, which can be defined as "ways of influencing choice without limiting the choice set or making alternatives appreciably more costly in terms of time, trouble, social sanctions, and so forth. They are called for because of flaws in individual decision-making, and they work by using those flaws" (Hausman *et al.* 2010, 124; Mongin *et al.* 2018).

The activity of nudging is based on the division of the mind's architecture into fast, parallel, automatic, associative, effortless psychological processes ("System 1") and slow, serial, controlled processes that require effort and are governed by rules ("System 2"). Based on this bipartition, one can then distinguish the nudges that exploit unintentional processes in System 1 from the nudges that instead enhance agents' reflexive and self-control abilities. For example, thrusts of the first type are interventions that exploit decision inertia and bias toward the status quo, whereby people tend not to make choices other than those they are accustomed to, or not to change the given situation, even when a change in the status quo could be beneficial. Choice architects can intervene by setting the most rational or most beneficial option, predicting that the agent will most likely tend not to change it. Other examples include exploiting the framing effect, that is, the disposition of agents to have different reactions to the same information when it is phrased in different ways, and the use of explicit imagery to make certain information more salient (think of the design of cigarette packages to make smokers more aware of the harms of smoking). Second-type nudges enable individuals to translate their intention into actual choices and actions and to avoid falling into the traps of the weakness of will. Agents can better understand information regarding specific products or situations to make more rational choices. "Educational nudges" do not exploit cognitive or decision-making limitations but enhance deliberative and executive skills. Generally, nudges do not coerce people to choose and act in a certain way, but they "guide" behavior, allowing agents to choose and act otherwise.

Nudges can be deployed in a paternalistic framework, when the choice architecture guides individuals' choices for the purposes of increasing their welfare, or in a non-paternalistic framework, when nudges guide individuals' choices to produce more externalities. The distinction is unclear, however, because deficient individual choices can also create negative externalities (Carlsson *et al.* 2021: 216), but I only consider non-paternalistic nudges since my focus is on BE. Nudges of this kind can be employed to induce actions respectful of significant community goods, such as reducing resource consumption or adopting ecologically compatible conduct (Carlsson *et al.* 2021; Wee *et al.* 2021; Santos

Silva 2022), or for their generic moral effect, insofar they increase altruistic and generally prosocial actions (Gråd *et al.* 2024; Valerio *et al.* 2021; Dimant *et al.* 2022). We will call these kinds of behavioral interventions "moral nudges".

Like proponents of behavioral ethics, the advocates of internal moral enhancements recognize the limitations of human nature as well: Persson and Savulescu (2012), for example, accurately describe human moral psychology as an evolutionary product adapted to very different environmental challenges than the current global ones and list many biases that need to be corrected. Walker claims that we can try to answer the question of why evil actions happen with great frequency by invoking views about humans' "defective natures", or the fact that "humans are innately evil is" (Walker 2009: 28-29). However, some proponents of MBE are skeptical about the efficacy of BE tools. For example, Persson and Savulescu note that nudges should be easy to avoid so that agents are genuinely free to choose otherwise; for this reason, behavioral tools such as nudges "are better suited to make us overcome backsliding on isolated occasions or to make us execute what we already think is best for us, or to make us decide between roughly equally balanced alternatives" (Persson *et al.* 2012: 79, footnote 2). The problem then lies in the effectiveness of these tools over time, which is not related to the ethical issues raised by nudging.

3. *The Automation Problem*

Recently, John Danaher (2019) argued for MEE's moral and political primacy to MBE. Danaher takes a broad definition of "moral enhancement": all interventions that enhance human moral judgment and behavior fall into this category, which includes "anything that develops morally-relevant emotions (such as trust or empathy), or virtues (such as courage and generosity), morally-relevant reasoning capacities (such as evidential assessment, impartiality and lack of prejudice), or improves individual moral actions (such as helping and caring for others)" (40) and distinguishes between internal (BME) and external moral enhancement (MEE). Danaher considers the effects on individual capabilities of a specific type of MEE, namely the use of electronic or AI devices that drive us towards the desired goal. For example, smartphone apps that nudge individuals toward money donations for charitable associations or other moral purposes, or a bracelet that gives you an electric shock when you do something morally or behaviorally inadequate (41-42). But he adds that these devices imply a philosophy of nudge, "which is influential in the design of many of the contemporary behavior change policies, apps, and devices" (41, 49). So, we can generalize his conclusion to embrace all behavioral interventions, even those which do not resort to smart devices and intelligent algorithms.

Danaher's central thesis is that the asymmetry between external and external enhancements should lead us to evaluate the former as morally more acceptable than the latter. Such an asymmetry implies rejecting the Ethical Parity Principle (in its weakest form) formulated by Neil Levy. In its weak version, the Ethical Parity Principle (EPP) holds that if we find compelling reasons for considering morally problematic interventions to modify the external environment, we must apply the same reasons to internal interventions. Unlike the strong version, the weak version of the EPP does not require us to accept the ontological thesis of the extended mind; that is, we need not assume that the domain of the mental extends outside the heads of individuals to include aspects of the external environment as well (Levy 2007: 60-64). According to Danaher, Levy's Ethical Parity Principle is undermined by three salient moral differences between internal processes and external devices (*memory integration*, *fungibility*, and *consciousness*). Memory is dynamic, while information stored in a notebook o in a mobile phone is static; a destroyed notebook can be easily replaced, and the user can form new memories or store new information if she lost pictures and files, while someone who got her hippocampus destroyed has a permanent disability in creating new long-term memories; finally, internal functioning is more integrated into conscious experience than the functioning of external props, and the same goes about internal memories and memories stored in external devices.

The differences identified by Danaher are relative to the *nature* of internal and external processes, functioning, and content, and this "ontological" divide has practical implications: internal devices produce *internal automaticity*, defined as "the control of behavior by not-immediately-conscious neural networks" (Danaher 2019: 49), while external devices initiate *external automation* processes. External automation has effects that hardly integrate with our perception and understanding of the world because it can easily bypass our conscious moral reasoning. On the contrary, internal automaticity does not undermine the deliberative process. So, what is troubling with MEE is its impact on what we can call "reflective capacities". Actually, Danaher does not use this term. Still, I prefer to speak of "reflective capacities" instead of "conscious moral reasoning" in order not to take a markedly rationalistic position in metaethics and moral psychology. Reflective capacities can also be compatible with a sentimentalist and deliberative conception of ethics, thus leaving open the metaethical question of how to precisely define these reflective powers.

According to Danaher (47), a problem of political legitimacy arises here: for a political decision to be legitimate, it must follow reliable procedures that exhibit outcome-independent virtues, as well as produce predictably desirable consequences. How policymakers intend to achieve a specific objective is *also* meaningful from this perspective. The proceduralist view introduces the need

to respect values such as transparency, participativeness, and comprehensibility. According to Danaher, external enhancements are incompatible with the proceduralist idea because they threaten these values, violating the central commitment of liberal democratic democracies, i.e., the commitment to treating citizens as moral *agents*, as subjects capable of actively relating to the moral problems they encounter in their lives (including political challenges). Bypassing reflective capacities, they turn targets into *passive* recipients, which are manipulated to have a particular desirable output (behavior that becomes more altruistic, more generous, more just, etc., thus conforming to specific moral standards). Danaher concludes that internal enhancements are preferable to the external enhancements.

Before analyzing Danaher's proposal, I would like to point out that, appearances to the contrary, this approach seems in line with at least one of Levy's remarks in introducing the EPP. He claims that a mere difference between internal and external cannot lead to a refutation of the externality thesis, but it should be taken as a confirmation. The point is that external props are attractive to the extent that they succeed in securing a more conspicuous cognitive, emotive, or motivational gain than that achieved by internal processes. For this reason, we can find attractive the ontological hypothesis of the extended mind (Levy 2007: 59-60). However, Danaher draws a radically different conclusion from this "pragmatic" approach: the fact that external tools are inefficient or even harmful when used to morally enhance people because they turn agents into passive recipients is a reason to oppose them, but the same reason does not apply to internal modifications. The automaticity produced by internal modification is more acceptable than automation, and so the EEP is unsound. Allhoff, Lin, and Steinberg (2011) argue an analogous line. They argue that the spatial location of the enhancement does not entail any moral difference because there is no reason to believe that incorporation is morally questionable. The example they chose is perfectly in line with Levy's equality principle. There is no difference *in kind* between "a neural implant [that] gives access to Google and the rest of the online world [and] using a laptop computer or Pocket PC to access the same" (204). The embedding nature of the former is not diriment to the extent that both carry "the same capabilities with us" (204). Although there is no moral difference in *kind*, there may be one in the *outcome*. According to Allhoff, Lin, and Steinberg, the moral symmetry between the two enhancements is assured if they are both effective in securing a particular capability; so, they suggest that we have to look for potential moral differences, not in the way we enhance an organism, but in the effects and in the impact on human capabilities of the enhancement we employ.

4. *The Automation Problem and the Role of Relational Freedom*

I strongly agree with Danaher on two points: (1) it's not the external or internal nature of an intervention that makes the moral difference, but its effects on agency; (2) automation poses a problem of political legitimacy. In fact, I would go even further with the latter: automation establishes a certain political relationship between the agent and the recipient of the intervention. Take some political authority that uses a BE tool to push people toward more altruistic choices; this intervention establishes a not-reciprocal relationship between the nudger and the nudgee.

A reciprocal relationship presupposes some basic expectations on the part of the people interacting, partially modeled on certain standards internal to that relationship. Thus, it is part of the regular interaction between human beings to expect that no one will be harmed by the other without a valid reason for doing so and that specific adverse reactions to a violation of this expectation are appropriate because the wrongdoer has betrayed the minimum threshold of trust that characterizes the relationship. Of course, the degree of trust or suspicion we may have toward others also depends on the context; situations that are less secure or in which we have little information may motivate a cautious attitude, just as interactions with people with whom we are more familiar may change the nature of mutual expectations. Trust and the reactions that follow its intentional violation form the core of a very specific type of relationship, one between people capable of reciprocity, that is precluded when dealing with very young children or people with severe mental illness. The interaction that takes place between moral agents presupposes the adoption of a dual attitude: a normative expectation of others' behavior and a willingness to treat the other as a "participant" in a reciprocal relationship. When this twofold attitude is not possible, relationships are marked by less reciprocity, to the extreme end of the spectrum where an "objective" attitude prevails. In such a case, the other is no longer considered a responsible person but someone to be "cared for", "managed", or "directed" (Strawson 1962). The shift from the participative standpoint to the objective one characterizes the nudge intervention. Nudgers suspend participatory attitudes and adopt an objectifying perspective toward the nudgee. She is a passive recipient, or at least she is treated as such.

One can reply that the nudger has a valid reason to adopt an objective attitude because the nudgee is in a condition of moral deficiency similar to that of an ill person. However, this kind of generalization fails to consider that the moral quality of behavioral intervention also hinges on the type of relationship we expect between those who possess political authority and those who are the recipients of policies. The endorsement of an objective attitude on the part

of nudgers depends crucially on the relationship between citizens and public decision-makers and the mutual expectations that structure these relationships. In this broader context, citizens are not agents to be respected but patients to be managed. As Hausman and Welch (2010: 134) put it:

> If a government is supposed to treat its citizens as agents who, within the limits that derive from the rights and interests of others, determine the direction of their own lives, then it should be reluctant to use means to influence them other than rational persuasion. Even if, as seems to us obviously the case, the decision-making abilities of citizens are flawed and might not be significantly diminished by concerted efforts to exploit these flaws, an organized effort to shape choices still appears to be a form of disrespectful social control.

Even if the use of moral nudges does not harm people's autonomy and well-being, it still impacts on the expectations that citizens in liberal democracies may have of those who govern them. The question of whether behavior management conflicts with such expectations is not merely empirical because it concerns the background against which interpersonal relationships are given; the introduction of nudges alters this background without this transformation being subject to reflection and consideration. This objection can also be framed in terms of respect for decisional autonomy (Rebonato 2012: 200-207). Still, regardless of the normative language it expresses, it echoes some criticism of internal moral enhancements. For example, Robert Sparrow (2014) pointed out a fundamental disanalogy between moral enhancement through traditional ways, such as education, and MBE: traditional means establish a relationship of equality and respect between the enhancing subject and the enhanced one, which responds to norms that justify educational activity and are in principle acceptable to all involved in the enterprise.

On the other hand, biomedical interventions create an entirely different relationship because they "operate in an instrumental or technical mode" (Sparrow 2014: 26) that treats the enhanced as an object and not as a subject. Sparrow echoes Philipp Pettit's idea of freedom as the absence of domination. In fact, the imposition of an MBE puts the enhanced person in a condition of subjugation and deprives her of her status as a responsible agent. Similarly, Michael Hauskeller (2017: 373-374) claims that if X makes it psychologically impossible for Y to want to do anything other than what X desires, then X's control over Y is total. Employing MBEs assumes an objectifying attitude: it expresses a suspension of all participatory perspectives and induces one to regard those who behave unjustly not as moral agents harboring inadequate moral dispositions or feelings while remaining fully participants in the practices of moral responsibility but as objects to be manipulated and corrected.

Thus, the same kind of criticism has been leveled at MBEs that target dispositions and emotions, as well as at MEEs that exploit automatic bias and heuristics. In both cases, the recipient of the intervention performs specific actions because of an automation mechanism: when this state of affairs is the intentional product of an institution or another person, a relationship of domination and control is established, and it is incompatible with the recipient's relational freedom. Moreover, MBEs and MEEs are also on a par regarding outsourcing moral reasoning. Danaher (2019: 50) stresses that external enhancements outsource the reasoning process, relieving agents of a cognitive burden, but the same issue seems to affect both MBEs and MEEs. In the case of internal enhancements, it is biochemical functioning that causes the output, bypassing reflective processes. If X receives a drug that amplifies his sense of justice in negotiations, her choices will automatically be more just without any reflective activity on his part. As Danaher claims, "We don't need to think for ourselves; we don't need to weigh the moral reasons for and against a particular action; the algorithm does all that for us" (50). If we substitute "algorithm" with "the molecular action", we have a mirrored criticism of MBEs.

Danaher has another arrow in his quiver: internal automaticity can easily be integrated with the individual mindset. A reiterated use of MBEs can be transformed into a permanent disposition in the long run and incorporated into one's moral character. For instance, a bioenhanced judge may take more empathic decisions without being aware of "the immediate proximate cause of his or her decision to choose the morally superior outcome, but he or she may over time generate a more empathetic disposition, which will affect future interactions with the world, and will, over time, result in enhanced moral sensitivity and awareness" (49). Even external enhancements rely on non-transparent mechanisms which the subject is not aware of. Still, they cannot be integrated in this way and can have corrosive effects in the long run (for a different view, see Agar 2014: 46-47).

However, even in this case, there is no real difference. It may be that the empathic disposition fits in the individual moral character in a more spontaneous way than the prolonged effect of an external factor can do; the judge may progressively endorse the effect of BME on dispositions. But note that if the judge does not voluntarily choose to undergo BME or has not had a pro-attitude toward it, his case looks very much like the manipulation of his moral character over time without giving any consent to these changes. In the case of a nudge, if some degree of transparency is assured, the agent is aware that his behavior can be directed in a certain way by an external prompt and willingly accepts the outcome of this conditioning. The automatic choice can become integrated into his identity. It seems that in both situations, the only relevant factor is the

degree of subjective awareness of the impact of MBEs or MEEs on judgment, choice, motivation, and behavior.

5. *Defeating the Ethical-Political Illegitimacy of Moral Enhancement*

To summarize, there are plausible reasons to deny that there is a morally relevant difference between BMEs and MEEs concerning the effects on moral agency. In both cases, the same kind of automation puts the enhanced or nudged individual under the control of who administers an enhancing drug or introduces a nudge in the choice context. In both cases, the same issue of political legitimacy arises. At this point, two conditions should avoid the concerns raised by automation. These conditions are defeaters that block the normative power of automation.

The first defeater is the presence of awareness or voluntary endorsement of the intervention. Alfano and Robichaud (2018: 242-244) see using (moral and non-moral) nudges as a responsibility-conferring practice. When institutions and private officers nudge individuals, they exercise power to attribute to the nudgees a forward-looking responsibility for distinct values. The values realized by nudges are varied; they claim that nudgers are more justified in using nudges when these tools induce individuals to fulfill obligations to themselves or others, while the power to resort to them is less supported when it is at stake the production of goods for self and others. A nudger has the ability to assign forward-looking responsibility for meeting some obligation via nudges only when the nudgee is liable for this assignment.

However, there are domains in which nudges are immune from the attribution of responsibility. For example, it seems morally unwarranted to nudge for sensitive choices such as marital decisions, voting, or healthcare decisions (although the introduction of nudges in the medical context is controversial and there are many proposals to use behavioral economics tools for clinical decisions or to obtain patients' informed consent more easily). Furthermore, they mention another possibility of being immune from nudge: individuals and communities can repudiate the responsibility assignment explicitly (i.e., through voting) or hypothetically (while it is unclear what form a hypothetical repudiation can take). Similarly, they can accept responsibility for values through an explicit or hypothetical endorsement, thus conferring political legitimacy to nudges (246). Alfano and Robichaud focus on the community and political levels. Still, an individual can reject a nudge if they are aware of its existence and operation, as I have already mentioned.

Further, citizens can become choice architects and opt for self-nudging (Reijula *et al.* 2022). Even in the case of BMEs, the agent can take a position by ac-

cepting or rejecting the intervention. This is the case of the so-called voluntary BME: one who freely chooses to bio-enhance herself shapes her future desires and intentions and expresses a solicitude for the moral quality of his future self. Voluntary BME is thus a strategy of preventive self-control, an "essential constraint" to use Jon Elster's terminology, i.e., a dodge by which agents self-impose restrictions to condition future behavior because of some expected benefit: voluntary BME expresses a "certain form of rationality over time" (Elster 2000).

The second defeating condition is related to a change in the target of biological or behavioral interventions. Such strategies can enhance reflective capacities instead of biases, emotions, and dispositions. The problem with voluntary and non-voluntary BMEs and MMEs is that they modify behavior, leaving the individual moral character untouched. They maximize good outcomes but do not correct moral flaws (Simkulet 2016). But we have other biological and behavioral ways to obtain real *moral improvement* in individuals.

Indeed, a possible alternative approach to nudging is the so-called "boosts", "interventions that make it easier for people to exercise their agency by fostering existing competencies or instilling new ones" (Hertwig *et al.* 2017: 974). The boosts approach shows significant differences from the nudge approach. First, it views agents not as passive recipients but as decision-makers "whose competencies can be improved by enriching [their] repertoire of skills and decision tools and/or by restructuring the environment such that existing skills and tools can be more effectively applied" (Grüne-Yanoff *et al.* 2016: 152). Second, it is interested not only in the *outcome* of decision-making (the conformity of behavior to specific standards of rationality and/or morality) but is concerned with the *process* through which such an outcome is achieved. Third, it does not demand to adapt the individual mind to the choice environment by exploiting its cognitive flaws in order to guide behavior but modifies the choice environment to suit the reflective powers of human beings. Fourth, its concern is in decision makers being aware of the limits of their minds and the errors they make in their judgments and decisions: the boosting approach requires the active cooperation of individuals (they are offers that can be accepted or declined).

Thus, the boosting approach aims to enhance subjects' cognitive, reflective, and deliberative features or, to use the language of the dual structure of the mind, they seek to educate System 1, by employing tools such as reminders, warnings, information labels, etc. The functioning of boosts is not dissimilar to the nudges that Sunstein calls "educational" or to other alternative approaches in BE as ethical debiasing, training, and moral disambiguation. Take, for example, the last one. In many situations, there is ambiguity about the existence of a conflict of interest, so that people tend to convince themselves of the absence

of the conflict and to excuse their immoral choices. A choice architecture that eliminates ambiguity can partly resolve the problem and mobilize individual moral resources to avoid immoral behavior (Feldman 2018: 98-100, 102-104).

Similarly, MBEs that succeed in directly amplifying or indirectly facilitating capacities of moral reflection seem to avoid automation. This is referred to in the literature as "procedural" or "indirect" MBE, which does not target specific moral dispositions, but enhances the capacity to correct feelings and instinctive reactions, drawing on a wide range of cognitive and noncognitive, individual, and social resources (Raus *et al.* 2014: 268-270; Schaefer 2015; Schaefer *et al.* 2019). Procedural MBEs do not guarantee effective behavior change, but they are, in principle, more respectful of individual moral agency. They allow the enhanced agents to make free choices and thus moral mistakes, hopefully learning from them. In addition, they allow "the enhancer to remain neutral on a wide range of substantive moral positions". Even if the enhancers "cannot be completely substantively neutral, […] the range and type of substantive issues within the scope of the enhancer are severely limited" (Schaefer 2015: 274). From a metaethical point of view, those who defend procedural MBEs cannot be neutral because they should take sides in the controversy between moral rationalism and moral sentimentalism. Nevertheless, the point is that enhancing moral deliberation, reasoning, and imagination can preserve moral agency from automation issues and the risk of being controlled by second parties.

Procedural MBEs and boosting MEEs can produce a moral *enhancement* that is simultaneously a real *moral improvement* because they make people more reflective in the moral domain without compelling them to make the morally correct choice.

6. *Conclusion*

Danaher's arguments against moral parity between MEEs and MBEs need to be more convincing because the risk of automation is substantial for both groups of interventions. But the distinction between automaticity and automation has heuristic value. It can serve as a basis for identifying finer-grained concepts for distinguishing enhancements that pose a problem of ethical-political legitimacy from enhancements that succeed instead in ensuring effective moral improvement. Moral boosts (or "educational nudges") and procedural or indirect forms of bio-enhancement fall into this second category and we have a moral reason to prioritize them over internal enhancements of moral dispositions and the use of nudges.

Finally, I would suggest that the internal or external location of ME interventions is not morally relevant; it is a different spatial metaphor that is morally

pertinent, namely whether the ME intervention is "high" because it targets individual reflective capacities, or "low" because it takes aim at automatic dispositions and behaviors.

Matteo Galletti
University of Florence
matteo.galletti@unifi.it

References

Agar, Nicholas, 2014, *Truly Human Enhancement: A Philosophical Defense of Limits*, The MIT Press, Cambridge.

Alfano, Mark, Philip Robichaud, 2018, "Nudges and Other Moral Technologies in the Context of Power: Assigning and Accepting Responsibility" in David Boonin, ed., *The Palgrave Handbook of Philosophy and Public Policy*, Palgrave Macmillan, Cham: 235-248.

Allhoff, Fritz, *et al.*, 2011, "Ethics of Human Enhancement: An Executive Summary", in *Science and Engineering Ethics*, 17, 2: 201-212.

Bazerman, Max *et al.*, 2012, "Behavioral Ethics: Toward a Deeper Understanding of Moral Judgment and Dishonesty" in *Annual Review of Law and Social Science*, 8: 85-104.

Borenstein, Jason *et al.*, 2016, "Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being" in *Science and Engineering Ethics*, 22, 1: 31-46.

Capraro, Valerio, Jagfeld, Glorianna, Klein, Rana *et al.*, 2019, "Increasing Altruistic and Cooperative Behaviour with Simple Moral Nudges" in *Scientific Reports* 9, art. N. 11880: 1-11.

Carlsson, Fredrik, *et al.*, 2021, "The Use of Green Nudges as an Environmental Policy Instrument", in *Review of Environmental Economics and Policy*, 15, 2: 216-237.

Danaher, John, 2019, "Why Internal Moral Enhancement Might Be Politically Better than External Moral Enhancement" in *Neuroethics*, 12, 1: 39-54.

DeGrazia, David, 2014, "Moral Enhancement, Freedom, and What We (Should) Value in Moral Behaviour" in *Journal of Medical Ethics*, 40, 6: 361-368.

Dimant Eugen *et al.*, 2022, "Meta-Nudging Honesty: Past, Present, and Future of the Research Frontier", in *Current Opinion in Psychology*, 47: 1-4.

Douglas, Thomas, 2008, "Moral Enhancement", in *Journal of Applied Philosophy*, 25, 3: 228-245.

Elster, Jon, 2000, *Ulysses Unbound. Studies in Rationality, Precommitment, and Constraints*, Cambridge University Press, Cambridge.

Eskine, Kendall J. *et al.*, 2011, "A Bad Taste in the Mouth: Gustatory Disgust Influences Moral Judgment", in *Psychological Science*, 22, 3: 295-299.

Feldman, Yuval, 2018, *The Law of Good People. Challenging States' Ability to Regulate Human Behavior*, Cambridge University Press, Cambridge.

Giubilini, Alberto *et al.*, 2018, "The Artificial Moral Advisor. The 'Ideal Observer' Meets Artificial Intelligence", in *Philosophy & Technology*, 31, 2: 169-188.

Gråd, Erik *et al.*, 2024, "Do Nudges Crowd Out Prosocial Behavior?", in *Behavioural Public Policy*, 8, 1: 107-120.

Grüne-Yanoff, Till *et al.*, 2016, "Nudge Versus Boost: How Coherent are Policy and Theory?", in *Minds & Machines*, 26, 1-2: 149-183.

Hauskeller, Michael, 2017, "Is It Desirable to Be Able to Do the Undesirable? Moral Bioenhancement and the Little Alex Problem" in *Cambridge Quarterly of Healthcare Ethics*, 26, 3: 365-376.

Hausman, Daniel M. *et al.*, 2010, "Debate: To Nudge or Not to Nudge", in *The Journal of Political Philosophy*, 18, 1: 123-136.

Hertwig, Ralph *et al.*, 2017, "Nudging and Boosting: Steering or Empowering Good Decisions" in *Perspectives on Psychological Science*, 12, 6: 973-986.

Lara, Francisco, 2021, "Why a Virtual Assistant for Moral Enhancement When We Could Have a Socrates?" in *Science and Engineering Ethics*, 27, 42: 1-27.

Levy, Neil, 2007, *Neuroethics: Challenges for the 21st Century*, Cambridge University Press, Cambridge.

Mongin, Philippe *et al.*, 2018, "Rethinking Nudge: Not One but Three Concepts" in *Behavioural Public Policy*, 2, 1: 107-124.

Persson, Ingmar *et al.*, 2012, *Unfit for the Future. The Need for Moral Enhancement*, Oxford University Press, Oxford.

Raus, Kasper *et al.*, 2014, "On Defining Moral Enhancement: A Clarificatory Taxonomy" in *Neuroethics*, 7: 263-273.

Rebonato, Riccardo, 2012, *Taking Liberties. A Critical Examination of Libertarian Paternalism*, Palgrave MacMillan, Basingstoke.

Reijula, Samuli *et al.*, 2002, "Self-nudging and the Citizen Choice Architect" in *Behavioural Public Policy*, 6, 1: 119-149.

Santos Silva, Marta, 2022, "Nudging and Other Behaviourally Based Policies as Enablers for Environmental Sustainability" in *Laws*, 11, 9: 1-13.

Schaefer, G. Owen, 2015, "Direct vs. Indirect Moral Enhancement" in *Kennedy Institute of Ethics Journal*, 25, 3: 261-289.

Schaefer, G. Owen *et al.*, 2019, "Procedural Moral Enhancement" in *Neuroethics*, 12: 73-84.

Schnall, Simone, *et al.*, 2008, "With a Clean Conscience: Cleanliness Reduces the Severity of Moral Judgments" in *Psychological Science*, 19, 12: 2008: 1219-1222.

Simkulet, William, 2016, "Intention and Moral Enhancement" in *Bioethics*, 30, 9: 714-720.

Sparrow, Robert, 2014, "Better Living Through Chemistry? A Reply to Savulescu and Persson on 'Moral Enhancement'" in *Journal of Applied Philosophy*, 31, 1: 23-32.

Strawson, Peter, 1962, "Freedom and Resentment" in *Proceedings of the British Academy*, 48: 187-211.

Walker, Mark, 2009, "Enhancing Genetic Virtue: A Project for Twenty-First Century Humanity?" in *Politics and the Life Sciences*, 28, 2: 27-47.

Wee, Siaw-Chui, *et al.*, 2021, "Can 'Nudging' Play a Role to Promote Pro-Environmental Behaviour?" in *Environmental Challenges*, 5: 1-13.

Wheatley, Thalia *et al.*, 2005, "Hypnotic Disgust Makes Moral Judgments More Severe", in *Psychological Science*, 16, 10: 780-784.

Creating capabilities to be better

Francesca Guma

Abstract: In this paper, I argue that the possibility of becoming better moral agents is related to the possibility of increasing both the opportunity and capacity to will otherwise and the effective conscious control of the will. Believing that it is essential for empirically informed ethics interested in moral enhancement to assess what to enhance and what type of enhancer is preferable, I begin by considering different types of enhancers and different factors on which they perform their action (§ 1). Secondly, I consider some issues arisen by moral enhancement in relation to the agent's freedom, emphasizing the need to reflect on the effects that such interventions can have on the agency (§ 2). I then propose a conception of free-will which can dialogue with empirical research in order to assess what the best moral enhancers might be and what factors they should act on to achieve real moral enhancements in individuals (§ 3). On such a basis, I assess what and how to enhance to achieve real moral improvement (§ 4) and present empirical proposals for procedural moral enhancement that leave open the possibility of achieving real individual moral improvement (§ 5). Finally, I conclude by stating that seeking out enhancers that can implement the opportunity and capacity to be good can lead to outcomes in which individuals do not become incapable of doing evil, but rather more capable of doing good (§ 6).

Keywords: moral enhancement, procedural moral enhancement, agency, free will, boost.

1. *Different types of moral enhancers, different factors to be enhanced*

In the contemporary ethical landscape, partly as a result of the considerations regarding self-control that psychologists, neuroscientists, philosophers, and decision theorists have made in recent years (Bermùdez 2018; Beeghly *et al.* 2020), several reflections and experimental proposals have arisen aimed at studying various aspects of moral reasoning to ascertain whether and how it is possible to enhance the ability to make moral choices (Klenk *et al.* 2021).

But is it possible to enhance the morality of human beings? If so, what needs to be enhanced to achieve real moral enhancement? What kind of intervention can be considered a good moral enhancer?

philinq XI, 2-2023, pp. 73-89
ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i2.493

In the lively debate on moral enhancement, it is possible to trace several proposals aimed at strengthening, reducing, and creating new capacities and/or dispositions of individuals in order to improve their moral actions. The approaches, although manifesting a common goal, are very different both in terms of metaethical and normative issues. The discussion of moral enhancement, in fact, involves elements that imply both normative aspects about the permissibility of certain interventions and an analytical discussion about the definition of morality and moral behavior (Reichlin 2019: 53). The purpose of this paper is not present and analyze all of these important elements, but rather to offer a normative reflection about what should be enhanced in order to achieve real moral enhancement of individuals and what – at least in principle – might be a good moral enhancer. To this end, I think it is appropriate to begin by focusing on just a few specific differences that can be detected by considering the most influential proposals in contemporary discussion.

Considering the different theories, it is easy to see that, on the one hand, the specific type of enhancer that is proposed, what we might call the active ingredient administered to the individuals to be enhanced, changes. For example, there are approaches that propose pharmacological interventions, others that advance technological strategies, and still others that propose more traditional interventions such as education. On the other hand, it varies what is intended to be enhanced, the factor on which the chosen intervention (the active ingredient) is to have its action. Some approaches, in fact, take into account motivation, others the decision-making process, and still others the outward behavior of individuals.

The different types of moral enhancers include those that act on the biochemical functioning of the subject and those that modify certain conditions external to the subject.

The first type is known as moral bio-enhancement, and argues for the possibility of improving the capabilities of human beings through interventions that act directly on the biological substrate (Persson *et al.* 2012). Making use of contemporary scientific and technical advances, such interventions are intended to act directly on biology to deliberately strengthen (or reduce) existing capacities and/or dispositions or create new ones, with the aim of improving an individual's behavior and moral capabilities. To achieve these results – at least in principle – some drugs could be administered, or genome editing interventions on the somatic or germ line could be implemented. When it comes to bio-enhancement, therefore, the goal is to use an enhancer that can modify a human being's behavior and judgment capabilities from within; in fact, in this view, it is proposed to change the original purpose of some technologies and drugs that were created to treat individuals with some disease in order to use them to take humans beyond the typical range of their natural abilities (Batistela *et al.* 2016; Maslen *et al.* 2014).

The second type of moral enhancers appears, at least in the first instance, to be less invasive because it does not alter the biology of the subjects, but rather aims to induce moral enhancements through external interventions of different kinds, such as through changing the context of choice, or increasing information, or using elements that influence sensory perception. These proposals include some that are more pioneering, investigating the possibility of harnessing advanced technological interventions such as artificial intelligence devices to make moral decisions (Giubilini *et al.* 2018; Lara 2021) and others that we may consider more traditional such as education, instruction, or nudging policies (Hausman *et al.* 2010).

Turning to the factor on which the moral enhancer (the active principle) is to perform its action, there are as many important differences that lead, in a general way, to divide the proposals into two groups: *substantive* (or direct) moral enhancement and *procedural* (or indirect) moral enhancement. The former are the interventions that claim to develop enhancers capable of acting on factors (biological, psychological, behavioral) capable of directly producing an effective change in the outward behavior of the enhanced individuals. These proposals aim to detect a change in the observable behavior or content of the normative judgments of the potentiated individuals (Schwitzgebel *et al.* 2020; Schwitzgebel *et al.* 2021). The second are the interventions that claim to use enhancers that act on factors (biological, psychological, behavioral) that can implement the moral capacities/dispositions of the enhanced subjects thus indirectly achieving improvements in moral acting (Schaefer *et al.* 2014; Schaefer 2015). These proposals aim at an improvement in the moral procedure of subjects, i.e., they are not directly concerned with the actual behavior of human beings, nor with the content of their moral judgments (although they consider both to be relevant), but rather aim to enhance the capabilities and faculties needed to ground them. What is relevant in the latter perspective is not what individuals do, judge, or believe, but rather the process that leads them to act, judge, or believe; in many cases we are faced with propositions interested in the reasons and justifications that can provide support for the actions, judgments, and beliefs of the individuals to be enhanced. This is no small difference, because on the one hand the focus is on the content of a given moral output, observable behavior or the content of normative judgments, while on the other hand the focus is on how a moral outcome is achieved, on seeking methods aimed at achieving an improvement in the skills and faculties involved in appropriate moral justice (Songhorian *et al.* 2022; Schaefer *et al.* 2019; Rawls 1951).

This distinction is as present in moral bio-enhancement proposals as in those that propose external interventions. In fact, the distinction between behavioral, emotional, and dispositional bio-empowerment is found in the literature (Jebari

2014). The former directly intervenes on the outward behavior, while emotional bio-enhancement and dispositional bio-enhancement (or combined forms of the two), on the other hand, aim to change the emotional or dispositional framework of individuals without directly changing their behavior. The latter types of interventions leverage cognitive and noncognitive resources and aim to improve the ability to correct feelings and instinctive reactions in order to achieve, precisely, indirect moral improvement. Such enhancers aim to enhance moral deliberation, or reasoning without acting directly to change the output in the enhanced subjects. Turning to enhancers that act from the outside to improve human morality, here again we find approaches that propose interventions that aim for substantive moral improvement, as in the case of nudging, and approaches that desire procedural moral improvement, as in the case of interventions that aim to implement education or knowledge of subjects (Songhorian *et al.* 2022; Schaefer *et al.* 2019; Schaefer 2015).

Regardless of whether a moral enhancer can act from within by modifying the biochemical functioning of individuals, or from without by modifying certain conditions in which individuals find themselves when acting or judging, it is crucial to consider what such an enhancer acts on and how it exerts its action in order to assess the effects it has on the subject's agency. The behavioral sciences are helping to provide a description of moral agency, emphasizing that the natural limitations of human beings require the use of innovative measures to help modify the actions and judgments of individuals through tools that can affect (more or less directly) decision-making. Achieving real improvement in the morality of individuals seems to depend at least in part on whether individuals can improve their own moral capacities, such as by reducing the influence of epistemic biases and other distorting influences. Based on empirical research, some believe that the pervasiveness of morally irrelevant influences on moral judgments prevents effective moral improvement (Klenk *et al.* 2021: 947-956). Other approaches, on the contrary, believe it is possible to achieve moral progress by improving our capacities to consciously control our moral judgments (Songhorian *et al.* 2022).

For empirically informed ethics it is not only crucial to understand whether and how moral enhancement is feasible, given the natural equipment of human beings, but also what to enhance and – at least in principle – what kind of enhancer is preferable and why. Believing that in this debate it is crucial to ask how much we can and should be satisfied with enhancing the outward behavior of individuals or whether we should aspire to a more arduous and ambitious goal of enhancing moral agency, I will now turn to consider some issues that moral enhancement raises in relation to the freedom of the subjects to whom the enhancement is to be administered. Our sense of agency is deeply tied to

the idea that we are capable of making a given choice independently when faced with at least two possible alternatives, which is why reflecting on the effects that moral enhancement has on freedom is important because it leads us to ask whether we are, and/or can remain, creators and masters of our own decisions and actions (Harris 2016).

2. *Moral enhancement and freedom: the importance of enhancing or granting agency*

One of the most recurrent critiques of moral enhancement alleges that it would compromise or eliminate the freedom of individuals. Such objections affect both freedom in its conception as free will and social freedom. Harris, for example, rejects certain forms of moral enhancement because they eliminate the autonomy necessary to be moral agents, not allowing the preservation of «freedom to fall» (Harris 2016). Sparrow, on the other hand, points out that certain moral enhancements violate social freedom because they treat individuals as an object, subjecting their mental states entirely to the will of others (Sparrow 2014; Hauskeller *et al.* 2018; Guma 2022).

Focusing on the issues that moral enhancement raises regarding free will, there emerges not so much a blind opposition to the possibility of tracing strategies aimed at moral growth, but rather a difficulty in accommodating proposals aimed at directly influencing the agent's outward behavior, regardless of the types of intervention.

This is the case, for example, with behavioral moral bio-enhancement, which by directly intervening in outward behavior is often considered to be the most invasive, coercive, and damaging form of personal freedom (Harris 2013; Harris 2011). Such enhancement deprives enhanced individuals of their status as responsible agents and places them in a condition of subjugation: enhanced individuals cannot do anything other than what is objectively considered good and, for this reason, are not considered true moral agents, but rather objects to be manipulated and corrected (Hauskeller 2017). Forms of emotional and dispositional bio-enhancement can also be subjected to an entirely similar critique, in that enhancing or eliminating certain emotions can still bring the subject into a coercive situation, making a certain response obligatory and necessary and eliminating the possibility of action alternatives. Being stressed strongly on an emotional level can lead to an incapacity to reflect and ponder, to choose and judge without evaluating the different factors that are important in that specific situation. For example, an individual with enhanced empathy or without antisocial emotions and self-interested motivations might feel such strong emotion

in perceiving the suffering of others that he or she has no alternative other than to act on the basis of that emotional reaction, without considering anything else. Even in such cases, therefore, the individual would not turn out to be free to choose otherwise, because he or she would not be able to contemplate any alternative (Songhorian 2019: 607-612).

If we shift to enhancements that act from the outside to improve human morality, when it comes to interventions that aim at substantive moral improvement the criticism remains the same: creating the external conditions capable of inducing subjects to engage in a specific behavior by bypassing the reflective process that would have led them to make that (or another) choice on their own eliminates (or at least reduces) the possibility of action alternatives. Nudges, for example, despite being presented as gentle nudges to influence choices without limiting or making possible alternatives particularly difficult (Thaler *et al.* 2008; Hausman *et al.* 2010; Mongin *et al.* 2018), in fact arise with the purpose of exploiting the psycho-cognitive limitations of human beings. By disregarding and circumventing the reflexive capacities of recipients, nudging does not hold individuals to be enhanced responsible agents, but rather "defective" people to be managed and directed; it treats subjects as objects to be manipulated to achieve a particular desirable outcome (Hausman *et al.* 2010). Leveraging biases and automatic heuristics leads subjects to perform specific actions because of an unconscious autonomation mechanism that is completely incompatible with the freedom of empowered subjects (Guma 2022: 190-193).

Using moral enhancers capable of acting directly on the behavior of subjects without them being able to notice it or, even worse, without them having any possible alternative in order to achieve effective change on the outward behavior of human beings does not appear to be a good way to improve their morality. Whether it is a bio-enhancer or an external enhancer, these are still interventions that aim to obtain a specific response from the enhanced individual, interventions that almost appear to be aimed at turning enhanced individuals into automatons under the control of those who administer an enhancing drug or introduce a nudge in the context of choice. The actions or choices made by individuals enhanced by this route do not appear to be what we are wont to regard as properly moral, precisely because they are not free actions or choices to which the individual gives assent after considering possible alternatives and assessing that a certain course of action is preferable. The reflection and weighing of possible courses of action, as well as the conscious control of one's choices and actions, are to be regarded as relevant, if not fundamental, aspects of morality. Inducing certain behaviors, eliminating or enhancing certain emotions and/or motivations, manipulating the reasoning process by making it at least difficult (if not sometimes impossible) for agents to bear the cognitive burden

of making a judgment or performing a certain action cannot lead the subject to be more moral. On the contrary, it seems to take away moral value from the action or judgment. Both in the case of bio-enhancement and in the case of some external enhancers aiming at substantial moral improvement, it is not properly the person considered who causes the action or makes the judgment, but some biochemical functioning or some mechanism of which the subject is unaware.

For these reasons, it is surely preferable to think of moral enhancement that, far from forcing us to certain actions that are reputed to be objectively better, allows us to broaden the range of considerations we are able to process. To aspire to a more ambitious project that can enhance people's agency – for example, by enhancing the ability to weigh reasons, or by improving the ability to provide justifications – seems like a way that is not only more respectful, but also closer to the real possibility of increasing the moral capacities of the subjects considered (Douglas 2014). This goal may – even rightly – appear very difficult to achieve (or in some cases even unattainable), but this does not legitimize the opportunity and desirability of enhancing mere outward obedience to certain moral standards. The latter path instead of making us more moral, may on the contrary make us less moral.

In contrast, some procedural moral enhancement proposals, in addition to being more respectful of individual freedom and moral agency, also appear to be able to achieve – at least in some cases – real moral improvement. Interventions that propose to use enhancers in order to achieve indirect improvements do not guarantee actual behavior change. These are proposals designed to enable enhanced individuals to make free and informed choices and to take actions that do not eliminate the possibility of making moral errors, but open up the possibility of noticing one's mistakes and learning from them. Moreover, a procedural account is preferable because it allows a pluralistic position to be maintained while avoiding substantive moral assumptions (Schaefer *et al.* 2019). Several moral controversies are, in fact, so problematic that it is difficult to believe that one solution is certainly the true one, which everyone has reasons to accept. Elaborating a procedural account that deals primarily with how people justify their behaviors, decisions, judgments, and beliefs is better suited to explain cases in which an individual might have come to a moral conclusion because of external or internal drives that would not be considered appropriate moral justification (Songhorian *et al.* 2022: 177-181).

Enhancers that aim to act on factors (biological, psychological, behavioral) capable of directly producing an actual change in the outward behavior of the enhanced subjects cannot be considered as good tools for achieving moral enhancement. Accordingly, proposals that advance the possibility of using enhancers to achieve substantial moral enhancement (regardless of the factor they

wish to enhance) cannot be accepted as good alternatives for achieving the goal. Claiming that individual moral improvement consists only in the performance of – or compliance with – practices judged by third parties to be morally better overlooks the possibility that behavior may be influenced or causally determined by drug taking, manipulation, or indoctrination. Such approaches fail to demonstrate whether the change achieved is brought about by actual, stable and genuine moral improvement precisely because it is not enough to observe an individual's behavior to understand it. People may act a certain way because they are influenced by internal or external stimuli, or, for example, by the desire to be socially approved. Based on the assumption that authentic moral action implies a strong sense of subjects' agency, focusing on seeking ways to enhance subjects' moral agency allows one to focus on what matters most in moral action, namely, how judgments are made and how moral behavior is grounded. In this way, the way can be opened to search for means and ways to increase the real moral capacity of agents and the conscientiousness of their moral responses.

In light of the above, I believe that reflection on moral enhancement must necessarily begin with an analysis of whether or not it is possible to implement the capability of agents to freely choose better actions and judgments. However, in order to understand whether there is real freedom to act or consciously choose something morally better, one must primarily give meaning to the expression "being free".

3. *An empirical conception of free will: the dynamic freedom of the will*

Considering the implications that moral empowerment may have on free will does not simplify reflection, because it adds a difficult topic of great theoretical but also practical importance to the issue. It is not my intention in this paper to offer an overview of the different positions and issues on this complicated topic (List 2019; Kane 2011). However, since I believe it is essential to take a position about free will in order to assess whether, what and how to enhance in order to achieve real moral enhancement, what I will do is to hint at some useful issues in order to understand the relevance of an investigation of free will to the discussion of moral enhancement and to support a conception of the freedom of the will that can dialogue with empirical research. In this way, it will then be possible for me to assess what the best moral enhancers might be and what factors they should act on to achieve real moral enhancements in individuals.

One of the reasons why I think it is essential to take a position on free will when it comes to moral empowerment is the fact that freedom of the will is usually interpreted as a condition of possibility for moral responsibility, in that

people are held morally responsible for their actions only if they could have acted otherwise: in a world in which freedom of the will does not exist, subjects cannot act differently from the way they have acted. How could we be held morally or legally responsible for an action if, in fact, that action does not depend on us? Moreover, from the perspective of subjects, free will is fundamental to the construction of their self-image as agents capable of deliberating over their own choices and actions. Our sense of agency is deeply linked to the idea that we are capable of making a given choice independently in the face of at least two possible alternatives.

In recent years, several scientists have argued that the concept of free will cannot be considered reliable scientific knowledge: people's choices are to be regarded as the product of neuronal activity in their brains, and it is unclear what role any intentional decisions should play. Many authors argue that thoughts and feelings are biochemical mechanisms that have little to do with an alleged freedom of will. Today, even following some psychophysical experiments, several researchers deny the existence of free will by appealing mainly to neuroscience, genetics, and cognitive science (Libet 2005; Harari 2016). If such considerations were correct, substantive enhancement could be seen in a less negative light: if we are not free to choose and decide, the use of tools that make us do and choose in ways determined by others cannot harm our freedom precisely because the freedom to will would only be an illusion. As much as several questions would remain open and unresolved (such as who can be elected as the correct and impartial decision-maker of the actions and judgments of all humankind, or whether or not it is appropriate to take a position of moral and cultural relativism that can take into account socio-cultural differences in the development of ethical norms or principles) in principle, substantive moral empowerment would no longer generate problems regarding the agency of the subject.

But theories denying free will are not supported by all. Some authors, taking ontological positions with regard to free will, accept the findings of contemporary research, agree on the existence of determinism but come to support compatibilist positions, concluding that although the will is determined, it is still possible to be free (Steward 2012). These positions do not support radical materialism because they disagree that, in a scientific worldview, it is necessary to take a reductionist, neuroscientific stance of human behavior (List 2019). It is not necessarily indispensable to focus on the biological and genetic constitution of the central nervous system for the purpose of analyzing phenomena such as free will; indeed, doing so may be entirely unnecessary or, in some cases, limiting. It is well possible that every mental event is related to specific electrochemical interactions occurring in neurons, that every mental event corresponds to a physicochemical event in the brain. This, however, is not enough to embrace

a reductionist thesis, because to understand the behavior of individuals it is not enough to observe their brains: not everything that is cerebral has psychic valence; and not everything that has psychic valence can be traced to the level of the brain. Observing physiological changes is not sufficient to understand the subject's psychic activity for one simple reason: to understand it, it is not possible to do without the subject and its psychological aspect, because to understand it, the aspect of sense and meaning is inescapable. This, of course, does not at all exclude the possibility (and necessity) of making neuroscientific investigations, or of seeking a correlation between the psychological and neurophysiological aspects.

Agreeing with these second positions, I believe it is possible to support a compatibilist and naturalistic theory of free will capable of pointing the way to concretely ascertain the actual presence or absence of the freedom of the will. Indeed, taking up Magni's concept of «*dynamic freedom of the will*» (Magni 2019: 62-74) it is possible to assume an empirical interpretation of the two conceptions deemed necessary to define free will: the existence of alternative possibilities and the conscious control of the will by the agent. According to this conception, people can be said to be free if they possess the opportunity and capability to act and will (i.e. if they have favorable external circumstances and internal conditions to act and will). In other words, one is free if one has no external impediments that hinder one from choosing and doing something and if one possesses psycho-physical characteristics that enable one to choose and do that something. This is an approach that distances itself from the classical metaphysical formulation of free will and considers freedom of will to be possible when one interprets it not in the sense of spontaneity, but as *opportunity* and *capability*.

Admitting that psychic determinism is evidence (just think of the fact that for human beings nothing is truly neutral and that, therefore, depending on the characteristics of the objects observed one is driven to produce unconscious/automatic inferences that condition the results of reasoning) and avoiding radical materialism tending to trace every aspect of the human mind to the brain, this theory allows one to assess whether one is free, or not, depending on the cases under consideration and assumes that different degrees of freedom may exist. It is not, therefore, a metaphysical question, because only in the concrete (case by case and relating to the particular will being considered) does it become possible to ascertain *whether* and *how much* the individual is *actually* endowed with the opportunity and capability to choose/act.

Rereading the requirement of the subject's conscious control over the will from the perspective of opportunity and capability brings to light the importance of the subjective feeling of agency, that is, the subject's perception that he or she is endowed with the opportunity and capacity to act and will:

an individual to be free must have both the mental capacities (internal requirement) and the opportunity (external requirement) to do and choose. In this view, if a moral enhancement takes away the agent's ability to choose whether to do good or evil, thus depriving him of an internal requirement of his free will, it impairs his freedom even if it does not affect other capacities or opportunities for choice and action.

Starting from this dynamic conception of freedom, it becomes possible not only to grasp the connection between being free and being able to make choices that we might call *better*, but also to evaluate *what* and *how* to enhance in order to achieve real moral improvement.

4. *Opportunity and capacity to be good*

Proposals that aim to directly influence the agent's outward behavior (regardless of whether they are interventions aimed, for example, at eliminating antisocial emotions or enhancing empathy through bio-enhancement, or suggesting the subject through the technique of nudging) at the very least inhibit the agent's reflective and rational capacities, achieving not real moral enhancement, but mere behavioral enhancement. These are proposals that reduce, rather than expand, the subject's possibilities for conscious choice and decision-making. In contrast, enhancing free will (in the definition I have advocated) could achieve real improvement in people's moral capacities, because it would make individuals freer to act and choose otherwise, to reflect about their decisions, to provide informed and thoughtful reasons.

An enhancer aimed at increasing freedom of will opens up the possibility of achieving effective improvement in moral capabilities, because it aims either to strengthen capacities that individuals already possess, or to increase their opportunities. Similar proposals aim to help individuals improve not a specific action, choice or behavior, but their general moral capabilities. Giving great weight both to the subjective feeling of agency and to the subject's decision-making process, and considering the possibility that choices are influenced by various natural and social forces, setting as a goal the increase of free will is equivalent to setting the goal of having individuals personally find the right kind of relationship between their preferences and action, freeing themselves from the conditions that preclude their freedom. In this way, individuals could obtain (or increase) both the concrete possibility of wanting otherwise, and conscious control over their own will, and thus arrive at a choice that we might call *better* (though not necessarily *better in a moral sense*).

Placing the focus not on the content of moral judgments developed by individuals (or on outward behavior), but on their capacity to develop such

judgments, to be aware of them, to justify them, to acknowledge them, to argue them, to provide good reasons in their defense gives the possibility of producing a moral enhancement that is at the same time a real moral improvement, because it makes people more reflective in the moral sphere without forcing them to make the morally correct choice. Maintaining the goal of increasing agency, one seeks not a change in outward behavior (of output), but a change in the procedure underlying moral reasoning skills, an increase in awareness of the reasons for one's moral judgments.

In light of the above, a moral enhancement intervention can achieve its goal by preferring enhancers aimed at achieving moral improvements through traditional external interventions such as education and instruction, or otherwise practices that always leave the agent with the ability to actually make a free choice or action. The action of the moral enhancer should be directed toward two goals. First: to enhance or increase in the subject the ability to choose and/or act otherwise, that is, the enhancer should seek to eliminate or diminish the effect of exogenous forces that constrain the individual from the outside (to help the subject have the actual opportunity to act and will) and the effect of endogenous forces that limit or prevent the individual from being able to act otherwise (to help the subject have the actual ability to act and will). Second, to enhance or increase the subject's effective conscious control of the will, that is, to enhance the subject's agency.

From this perspective, substantial moral enhancement is to be rejected, regardless of whether it is achieved through pharmacological, technical interventions or those aimed at modifying certain conditions external to the subject. The rationale lies not only in the fact that such interventions do not respect the freedom of the subject, but also in the fact that they would not produce real enhancement of the moral capabilities of the enhanced individuals. On the contrary, procedural moral enhancements can provide us with strategies that – at least in principle – can actually improve and/or enhance capacities to make moral choices. In the next section I mention some empirical proposals that seem likely to achieve this goal.

5. *Empirical proposal to be freer and, perhaps, better*

Assuming the importance of rational awareness in the formulation of moral judgments and affirming the ineradicable presence of endogenous and exogenous conditions that make full self-control of action and choice difficult, individuals are not, however, precluded from moral improvement: such improvement can be achieved by increasing their effective possession of the opportunity and capability for conscious control.

The project is ambitious, but there are investigations and reflections that make it seem feasible. For example, recent studies show that in some contexts our implicit biases can also be changed easily (Johnson 2020), while some authors propose strategies capable of having people develop indirect control over such biases, for example through the development of a set of long-term habits or certain social policies (Beeghly *et al.* 2020: part III). These are interesting proposals, which should be investigated further to assess whether they can really contribute to the development of procedural moral enhancements that can improve the opportunity and capability to be better by leaving the individual free to act and choose.

It is also useful to consider the problem of adaptive preferences highlighted by Elster and Sen: what individuals prefer responds to the social and environmental conditions in which they are embedded, which is why their choices may often not be what they would make if they were more aware of their situation (Elster 1983; Sen 2009). Sometimes, by increasing information, individuals act differently than they would have done by ignoring certain factors. An increase in knowledge may lead not only to an increase in the opportunity to act otherwise (precisely because one may come to know alternatives that one did not know before), but also to an increase in awareness, because it leads one to re-evaluate, reflect on, and weigh factors that one did not know.

Another interesting proposal, and not far from these considerations, is that of Gigerenzer, who stresses the possibility and importance of educating individuals to make the best possible decisions independently. Another interesting proposal, and not far from these considerations, is that of Gigerenzer, who stresses the possibility and importance of educating individuals to make the best possible decisions independently. Confirming the impossibility of «Olympic rationality», Gigerenzer points out that intuitive, quick, and immediate mental processes are useful, often necessary, and capable of leading to optimal decisions if one has the proper tools and knowledge to avoid falling victim to bias and the way information is presented (Gigerenzer 2008). His proposal can also be read from the perspective of developing interventions aimed at boosting the subject's agency. It is, in fact, a boosting proposal, that is, an intervention that makes it easier for people to exercise their agency because it aims to promote existing skills or, where appropriate, instill new ones (Hertwig *et al.* 2017). Theories supporting boosts view agents as active recipients of boosting interventions, deciders whose skills can be enhanced by enriching their repertoire of decision-making skills and tools. The idea is to boost the skills that individuals already possess, or to modify the external environment to allow existing skills and tools to be applied more effectively (Grüne-Yanoff *et al.* 2016). The boosting approach aims to improve subjects' cognitive, reflective, and deliberative

characteristics, requires their active cooperation, and proposes aids or reinforcements that can be accepted or rejected. For this reason, it can be considered as another good empirical proposal for achieving real moral improvements.

Further proposals, based precisely on the assumption that authentic moral action involves a strong sense of agency on the part of individuals, advance theories of procedural moral enhancement that leverage the importance of finding strategies that focus on how judgments are made and how moral behavior is grounded. Having as their goal the enhancement of justification to moral actions and judgments, such approaches seek methods that focus on individuals' abilities to provide reasons based on logical, empirical, and conceptual competence, openness to revision of one's opinions, sympathetic imagination, and reduction of biases (Songhorian *et al.* 2022: 179-181).

Making information more understandable, educating to recognize one's own limitations and mistakes, increasing the possibility of receiving feedback, providing spaces for discussion, developing *ad hoc* education programs, facilitating increased awareness, and helping to improve one's capability to map decisions can be considered some of the ways to increase the positive aspect of the individual's free will and stimulate moral growth. Certainly, such strategies do not generate automatons with perfect morality, however, they leave open the possibility of achieving real individual moral improvement.

6. *Conclusion*

Devising procedures that strengthen the free will of the subject makes it possible to think of genuine and stable moral improvements because these are interventions aimed at achieving changes and improvements not in specific outward behaviors, but in the general and natural moral capabilities of individuals. From this perspective, the risk of possible coercive effects in the area of social freedom is also avoided: by focusing not on the content of judgments but on the way in which they are made, no specific normative framework is assumed, nor is it assumed to list what is good or right, leaving agents free to develop the judgment they deem most appropriate, thus avoiding interference or domination by others.

Considering that individuals' sense of agency is intimately connected to the idea of being able to make decisions autonomously and consciously, keeping agency enhancement as a goal safeguards a morally relevant characteristic: as much as the literature on moral responsibility provides different perspectives about what makes a subject responsible for an action, it is quite common to believe that being morally responsible for an action has a deep connection with what it takes for that action to be an expression of the agent's will. Therefore,

thinking of ways to increase people's agency seems a good way both to increase their moral capacity and to respect them (Reichlin 2017).

Considering a naturalistic conception of free will in relation to a theory of moral enhancement makes one reflect on the need to think about interventions that help the individual increase his opportunities and capability to act, will, and choose consciously when exercising his moral capacities. It also highlights the importance of the agent reinforcing the power or capability to judge something good or right, without having external judges deciding for him.

All this does not exclude the possibility of educating subjects, of trying to convince them to make some choices rather than others, but it does not admit the possibility of obtaining automatons with perfect morality. Searching for enhancers who can implement the opportunity and capability to be good, conceiving of moral enhancement as indirect, from a formal rather than a content-based perspective, can lead to seeing realized situations in which individuals do not become incapable of doing evil, but rather potentially more capable of doing good.

Francesca Guma
Vita-Salute San Raffaele University
guma.francesca@hsr.it

References

Batistela, Silmara *et al.*, 2016, "Methylphenidate as a Cognitive Enhancer in Healthy Young People" in *Dementia & Neuropsychologia*, X, 2: 134-142.

Beeghly, Erin *et al.*, 2020, eds., *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*, Routledge, New York.

Bermùdez, José Luis, 2018, ed., *Self-Control, Decision Theory, and Rationality: New Essays*, Cambridge University Press, Cambridge.

Douglas, Thomas, 2014, "Enhancing Moral Conformity and Enhancing Moral Worth"in *Neuroethics*, VII: 75-91.

Elster, Jon, 1983, *Sour Grapes: Studies in the Subversion of Rationality*, Cambridge University Press, Cambridge.

Gigerenzer, Gerd, 2008, *Rationality for Mortals: How People Cope with Uncertainty*, Oxford University Press, New York.

Giubilini, Alberto *et al.*, 2018, "The Artificial Moral Advisor. The 'Ideal Observer' Meets Artificial Intelligence" in *Philosophy & Technology*, 31, 2: 169-188.

Grüne-Yanoff *et al.*, 2016, "Nudge Versus Boost: How Coherent are Policy and Theory?", in *Minds & Machines*, 26, 1-2: 149-183.

Guma, Francesca, 2022, "Becoming Better Moral Agents by Strengthening Free Will.

A Possible Prospect?" in *Teoria. Rivista di Filosofia*, 2: 187-199.

Harari, Yuval Noah, 2016, "Yuval Noah Harari on Big Data, Google and the End of Free Will" in *Financial Times*, 26 August 2016. Harris, John, 2011, "Moral Enhancement and Freedom" in *Bioethics*, XXV, 2: 102-111.

Harris, John, 2013, "Moral Progress and Moral Enhancement" in *Bioethics*, XXVII, 5: 285-290.

Harris, John, 2016, *How to be Good*, Oxford University Press, New York.

Hauskeller, Michael, 2017, "Is It Desirable to Be Able to Do the Undesirable? Moral Bioenhancement and the Little Alex Problem", in *Cambridge Quarterly of Healthcare Ethics*, 26, 3: 365-376.

Hauskeller, Michael, *et al*., 2018, eds., *Moral Enhancement: Critical Perspectives*, Cambridge University Press, Cambridge.

Hausman, Daniel M., *et al*., 2010, "Debate: To Nudge or Not to Nudge" in *The Journal of Political Philosophy*, 18, 1: 123-136.

Hertwig, Ralph, *et al*., 2017, "Nudging and Boosting: Steering or Empowering Good Decisions" in *Perspectives on Psychological Science*, 12, 6: 973-986.

Jebari, Karim, 2014, "What to Enhance: Behaviour, Emotion or Disposition?" in *Neuroethics*, VII, 3: 253-261.

Johnson, Gabbrielle M., 2020, "The Psychology of Bias: From Data to Theory" in Editors Erin Beeghly, Alex Madva, eds., *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*, Routledge, New York: 20-40.

Kane, Robert, 2011, ed., *The Oxford Handbook of Free Will*, Oxford University Press, Oxford.

Klenk, Michael *et al*., 2021, "Moral Judgement and Moral Progress: The Problem of Cognitive Control", in *Philosophical Psychology*, 34, 7: 938-961.

Lara, Francisco, 2021, "Why a Virtual Assistant for Moral Enhancement When We Could Have a Socrates?" in *Science and Engineering Ethics*, 27, 42: 1-27.

Libet, Benjamin, 2005, *Mind Time: The Temporal Factor in Consciousness*, Harvard University Press, Cambridge.

List, Christian, 2019, *Why Free Will Is Real*, Harvard University Press, Cambridge.

Magni, Sergio Filippo, 2019, *L'etica tra genetica e neuroscienze. Libero arbitrio, responsabilità, generazione*, Carocci, Roma.

Maslen, Hannah *et al*., 2014, "Pharmacological Cognitive Enhancement-How Neuroscientific Research Could Advance Ethical Debate" in *Frontiers in Systems Neuroscience*, VIII, 107: 1-12.

Persson, Ingmar, *et al*., 2012, *Unfit for the Future: The Need for Moral Enhancement*, Oxford University Press, Oxford.

Rawls, John, 1951, "Outline of a Decision Procedure for Ethics" in *The Philosophical Review*, 60, 2: 177-197.

Reichlin, Massimo, 2017, "The Moral Agency Argument Against Moral Bioenhancement" in *Topoi*, XXXVIII, 1: 53-62.

Schaefer, Owen G., 2015, "Direct vs. Indirect Moral Enhancement" in *Kennedy Institute of Ethics Journal*, 25, 3: 261–289.

— *et al.*, 2014, "Autonomy and Enhancement" in *Neuroethics*, 7: 123-136.

— *et al.*, 2019, "Procedural Moral Enhancement" in *Neuroethics*, 12, 1: 73–84.

Schwitzgebel, Eric *et al.*, 2020, "Do Ethics Classes Influence Student Behavior? Case Study: Teaching the Ethics of Eating Meat" in *Cognition*, 203: 104397.

Schwitzgebel, Eric *et al.*, 2021, "Students Eat Less Meat After Studying Meat Ethics"in *Review of Philosophy and Psychology*.

Sen, Amartya Kumar, 2009, *The Idea of Justice*, Penguin Books, London.

Songhorian, S., *et al.*, 2022, "Moral Progress: Just a Matter of Behavior?" in *Teoria. Rivista di Filosofia*, 2: 175-187.

Songhorian, Sarah, 2019, "Che cosa conta come un (effettivo) potenziamento morale?"in *Bioetica. Rivista interdisciplinare*, 27, 4: 597-615.

Sparrow, Robert, 2014, "Better Living Through Chemistry? A Reply to Savulescu and Persson on 'Moral Enhancement'" in *Journal of Applied Philosophy*, 31,1: 23-32.

Steward, Helen, 2012, *A Metaphysics for Freedom*, Oxford University Press, Oxford.

Thaler, Richard H., *et al.*, 2008, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press, Yale.

Public Reason and Biotechnological Moral Enhancement of Criminal Offenders*

Elvio Baccarini

Abstract: There are two prominent classes of arguments in the debate on mandatory biotechnological moral enhancement (MBME) of criminal offenders. Some maintain that these interventions are not permissible because they do not respect some evaluative standards (my illustration is represented by autonomy). Others, however, argue that this type of intervention is legitimate. One of the latter argumentative lines appeals to the reduction of the high costs of incarceration. In this paper, I argue that such polarization in the debate suggests handling the problem of the protection of autonomy in the case of MBME of offenders as an allocative question. Moreover, I offer a novel approach to this question by adopting the Rawlsian method of public reason. According to this method, public decisions are legitimate only if they can be justified through reasons that can be accepted by each free, equal, and epistemically reasonable agent. I argue that, within this framework, for a specific class of criminal offenders, we can conclude that MBME, although undermining a certain form of autonomy, could be legitimately mandatory. Because of reasonable pluralism, the final verdict on legitimacy is made based on the results of fair procedures of decision-making among proposals supported by persons in a condition of reasonable disagreement.

Keywords: autonomy, allocative justice, biotechnological moral enhancement, public reason, Rawls.

1. *Introduction*

Scientific advancements and hypotheses concerning the modification of morally relevant behavior through technological resources have, in recent years, prompted discussions on the biotechnological moral enhancement (BME) (Harris 2016; Persson *et al.* 2012). Within these, the case of a mandatory biotechnological moral enhancement (MBME) of criminal offenders that refuse rehabilitation holds a number of prominent and promising discussions (Birks

* This paper is an outcome of the Research Project JOPS (Public justification and Capability Pluralism), financed by the Croatian Science Foundation (HRZZ) (PI Elvio Baccarini; grant number: IP-2020-02-8073). Initial work on the paper was financed as well by the Croatian Science Foundation (Project RAD - Responding to antisocial personalities in a democratic society IP-2018-01-3518; PI Luca Malatesti).

ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i2.494

et al. 2018). Broadly speaking, this type of treatment involves hindering antisocial behavior by modification of emotional responses, dispositions, or motivations, or directly impeding some behaviors of targeted individuals through biotechnological means. Such means are different from external coercion, like imprisonment, because they involve modifying a person's psychological, physiological, or mental structure.

Debates on BME most frequently refer to medical interventions, but my discussion could be applied to possible cases of moral enhancement through AI, as well as other resources (Savulescu *et al.* 2015; Giubilini *et al.* 2018). Some BME interventions are already applied in criminal justice, particularly in drug addiction and sex offending. Research is being conducted on the possibility of interventions to modify impulse control and reduce aggression (Ryberg 2012: 231; Shaw 2018: 251). Yet, the current resources for BME at our disposal are incomplete, to say the least. Thus, the possible immediate contribution of the paper, in practice, is not that of offering recommendations on whether to apply MBME. However, some authors remark that promising research is being conducted (Chew *et al.* 2018). This is enough to render the discussion in the present paper meaningful. Primarily, the intention is to contribute to the inquiry into whether it is morally permissible or even recommendable to invest intellectual and material means into attaining possible technological resources for realizing the MBME of criminal offenders resistant to rehabilitation. If such MBME could not be justified through valid moral reasons, a positive answer is excluded. Contributions of the present paper could be, for example, those indicating some principled reasons as relevant in the dispute about MBME and the proper method of deliberation. Such recommendations could be addressed, for example, to bodies that establish whether to sustain research of BME technologies with public funds, or not, to permit them, or to ban them.

Several rights and values are discussed in this context (Bublitz 2018; Douglas 2014; 2018; Shaw 2018; 2019). Disputes include theories of punishment as well (Matravers 2018; Ryberg 2012). An important line of argument in favor of MBME has been inaugurated by Thomas Douglas, a pioneer in the debate. He suggests that MBME is morally analogous to some accepted or intuitively acceptable practices, like incarceration or environmental modulation, which defuses criticisms that appeal to physical and mental integrity (Douglas 2014; 2018). I focus on a particular question here, and I try to formulate a plausible answer by employing a method of public justification originally present in theories of political philosophy. The method will be utilized to assess normative reasons employed by actual participants in the MBME dispute. Precisely, in the present paper, I focus on one of these reasons, the normative importance of autonomy – does it defeat the legitimacy of MBME in the case of criminal offenders?

I use 'autonomy' in the present paper, with two meanings. First, 'autonomy' corresponds to personal autonomy. It includes the capacity to choose actions through rational and free choices (Harris 2016: 78), as well as "the ability to determine for oneself the nature of the moral reasons, considerations and principles on which to act" (Bullock 2018: 162)[1]. Further, 'personal autonomy' refers to the capacity to do, or not to do, something (Harris 2016: 78), as well as to the capacity to act on the moral reasons, considerations, and principles that one has determined (Bullock 2018: 165-166). Personal autonomy includes the internal capacity of an agent. In other words, the capacities that a person has in virtue of her inner mental and physical abilities.

The other conception of autonomy that I discuss is autonomy as non-domination. Hauskeller endorses this type of view. For this view, the concept of autonomy is characterized socially. What comes to the fore is not our inner ability to make choices but the condition of not being dominated by others, i.e., being the authors of our lives in a social context. The domination complaint points to a different aspect of autonomy loss than the one shown above: the agent is bypassed, and others use their position of power to make decisions instead of her (Hauskeller 2017).

I argue that certain forms of MBME of certain criminal offenders could be legitimate under certain conditions. Precisely, I argue that MBME could be part of an eligible set of public decisions that could be the matter of choice in a fair procedure. My argument relies on two principal claims. First, following Matt Matravers (2018) and Jeff McMahan (2018), I maintain that the economic burden placed on society by incarcerating offenders is a relevant factor in the discussion of MBME. As such, the dispute about the MBME of offenders connects with issues of allocative justice (just allocation of resources).

Second, I employ and adequately adapt to the present debate, a general method of justification of public decisions. The method that I use is represented by John Rawls's theory of public reason (2005). This method requires that relevant public decisions are justified through reasons that all qualified persons can accept. In applying this method, within the present context, I assess the reasons of authors who participate in the debate on MBME. Thus, as defined previously in this discussion, I examine the prospect of an economic argument for the MBME of criminal offenders from a moral viewpoint, and I evaluate autonomy because authors who participate in the dispute are concerned about its normative strength. I discuss whether, under what conditions, and in what limits, this concept's employment is legitimate and what conclusions we can

[1] My terminology differs from that of some other authors. For example, Bublitz "denotes the psychological faculty to form and revise moral beliefs" by the term 'conscience' (Bublitz 2018: 300).

derive from its use. Finally, I analyze which conclusions are eligible.

Rawls's theory is complex and constituted by various components, such as, for example, his theory of justice and his theory of public justification. Although I am concerned with applying Rawls's method of public justification, I do not try to show what kind of conclusion for MBME follows from Rawls's theory of justice, nor am I committed to endorsing all of its elements. Instead, my goal is to assess the present debate and the use of reasons employed by actual participants through Rawls's method of public justification.

I see the employment of Rawls's public reason as important for a general question of political philosophy. His influential political-philosophical opus has been charged because of the alleged inability to deal with some crucial justice questions (Nussbaum 2006). In the present paper, I extend a part of his doctrine, precisely, his theory of public justification, to a domain sidestepped by him. In this way, I try to show that at least a part of Rawls's doctrine can be extended to embrace a wider domain of issues than those addressed by the author. Divergences that regard MBME represent a relevant issue for this method's employment. I explain why, although Rawls never discusses the current topic.

I will proceed as follows. First, I will clarify the exact scope of the discussion, the nature of the interventions discussed by my argument, and the types of offenders targeted. This part will explain the extent of my main conclusion.

In the second part, I present some reasons for MBME shown by Matravers (2018) and McMahan (2018). Such reasons point to the fact that considerations of costs could matter for the moral assessment of MBME.

In the third part, I describe arguments against the MBME of criminal offenders that represent challenges to the argument that economic reasons matter and that they could speak in favor of this practice. There is a variety of such arguments. Still, the focus of my discussion is the worry that the MBME of criminal offenders undermines autonomy.

In the fourth part, I illustrate the method of justification of public decisions that I employ in this paper.

Finally, I develop my central argument for my main thesis. The conclusion I derive from my application of the method of public reason is that there is not a unique reasonable decision whether we can (or, even, must) adopt MBME, or that it is inadmissible. We are not able to establish through valid public reasons a single decision as to the most reasonable. Instead, we arrive at a set of reasonable and, thus, eligible proposals. Consequently, final legitimate decisions on whether to opt for MBME or incarceration are derived from fair procedures. According to such procedures, agents decide between competing eligible proposals supported by valid reasons – i.e., reasons that are epistemically reasonable and that they can accept as free and equal citizens.

2. *Setting the debate*

Because I do not discuss the already available technologies, I cannot offer a detailed description of the BME interventions. In my discussion, I refer to all prospective interventions that could achieve BME as defined at the beginning of the paper. From a moral point of view, we can distinguish between two broad forms of BME: those interventions that modify aspects of the subject's personality (for example, their emotions or dispositions), and those that change only capacities for behavior (by making specific actions physically impossible or very unlikely, like causing nausea when an agent wants to perform a violent act). The latter could resemble ordinary social reactions to people who represent a threat to society or some criminal offenders, like incarceration. However, MBME is peculiar because it consists of a modification of the internal abilities, of a person's constitution, and not only of external limits to the capacities to act. In MBME, we change how a person is and who they are (although not always so profoundly as when we change a person's character traits). In other words, we do not change only what the person can or cannot do. This is why MBME can be a problem, even if external limits to people's behavior, like incarceration, are accepted. The problem is present because, at least in usual liberal views, the state's jurisdiction includes regulation of persons' behavior according to justice requirements. That is, the state stands limited to change the persons' internal characteristics (Jacobs 2016).

Despite arguments by authors who find interventions that modify aspects of the subject's personality worse than interventions that change only capacities for behavior (Shaw 2011: 197, 200-202), my argument aims to show the possible legitimacy of both kinds of interventions. I include all such cases under the label of 'direct interventions' as I refer to modifications that are not consciously mediated by their subjects. For instance, those represented by electrochemical or physical reactions (Bublitz *et al.* 2014: 69). Distinct from such direct forms of moral enhancement is an indirect enhancement, accomplished, for example, through powerful rhetorical speech (Liberto 2018: 203), or through cognitive enhancement that leads to better reflection, for example, by reducing distractions and impulsive reactions (Bullock 2018: 167)[2].

[2] To be sure, I concede, but do not endorse the conceptual distinction between direct and indirect interventions. To illustrate, Kasper Lippert-Rasmussen's arguments that deny the difference between direct and indirect interventions (2018), or the doubts raised by Hallie Liberto (2018: 203), are highly relevant. But my intention is to show that there is a reasonable argument in favour of the more controversial of the two kinds of interventions, i.e. mandatory interventions that are clearly direct, and, thus, I do not need to take a stance on this question. A consequence of the conceptual concession might be that indirect attempts of modifications for the ends of pro-socialisation and rehabilitation are to be favoured over direct modifications, and the latter can be legitimate only after the former have persistently failed.

I limit the discussion to criminal offenders that have committed particularly grave crimes, who oppose traditional forms of rehabilitation, and are a persistent clear and present danger to society. I do not defend the extension of MBME to all criminal offenders, nor do I rule out that such a justification can be offered. An obvious problem, at this point, is how can we determine whether some convicted criminal offenders are, in fact, clearly present and persistent dangers to society (Matravers 2018: 85). This is a relevant question, and uncertain cases require a separate discussion, which is vital for the possibility of extending MBME. This would be important for gaining more comprehensive benefits from MBME. However, I skip this question in the present paper. Instead, I shape the illustrations of my discussion's scope to clarify that we have persons who represent a persistent and clear danger to society. This is because I want to focus on the fundamental question addressed by opponents of MBME discussed in the present paper. Are there principled reasons, evaluative standards, that outweigh the possible benefits of MBME of criminal offenders who are resistant to rehabilitation even in the most extreme cases? If the answer is affirmative, the discussion on MBME would be resolved at the very beginning.

One of the two case illustrations discussed in the present paper uses the film character Hannibal Lecter. The other is a politically radical terrorist. Let us call him Anders. Anders has committed politically motivated mass murder. He persists in affirming the rightness of his criminal act, the ideology that had motivated it, and the readiness to repeat his actions. Imagine that Hannibal Lecter is exactly like the character in *The Silence of the Lambs* but does not suffer any pathology that impairs him in making choices. He is deliberately a cruel person who takes pleasure in other people's suffering and humiliation and has, in the background, adopted an intellectual construction that justifies such an attitude to himself. He is delighted with his motivations, and they fully correspond to his evaluations. Like Anders, he wants to continue being like he is and repeat the same crimes if given the opportunity. The two persons clearly have unacceptable conceptions of justice and the good. There are no doubts that society can legitimately impede them from committing violent criminal acts that harm other persons. But the moral problem for MBME is represented by the tendency of liberal states to use coercion only for regulating external behavior, and not for changing inner conditions of persons, like, for example, particular dispositions, character traits and internal abilities to act. Such conditions are not in the domain of

This holds if we can prove that indirect interventions are preferable to direct intervention (which I concede). But I interpret this thesis as depending on what I only take to be a conditional concession, not an endorsement of the distinction between direct and indirect interventions.

the state's legitimate use of coercion (Jacobs 2016).

The scope of the application of MBME that I defend in the paper partly overlaps with Ingmar Persson, and Julian Savulescu's God Machine thought experiment. The God Machine is an imaginary device that controls people's behavior to exclude the worst forms of antisocial behavior and related character traits. I refer to this heuristic device to indicate the scope of moral enhancement that I have in mind to be achieved via prospective biotechnological means. This is how Persson and Savulescu describe it:

> The God Machine was designed to give human beings near-complete freedom. It only ever intervened in human action to prevent great harm, injustice, or other deeply immoral behavior from occurring. For example, the murder of innocent people no longer occurred. As soon as a person formed the intention to murder, and it became inevitable that this person would act to kill, the God Machine would intervene. The would-be murderer would 'change his mind.' The God Machine would not intervene in trivial immoral acts (Savulescu *et al.* 2012: 411).

The range of interventions of the God Machine example does not overlap precisely with instances of MBME that I discuss in the present paper. In Persson and Savulescu's proposal, it is applied to all agents to pre-empt the morally worst actions. In the cases that I discuss, MBME is used only after an agent has already committed a horrible crime and the person persistently refuses rehabilitation via traditional means. Furthermore, in the cases that I discuss, the person is already deprived of freedom. However, the range of interventions of the God Machine and the cases that I discuss overlap in that, in both situations, the worst and only the worst moral actions are impeded.

I limit the discussion to agents who are not diagnosed with any condition that would restrict their autonomy. I presuppose that these agents deliberately choose criminal behaviors. This is because the relevant normative constraint highlighted by opponents of MBME that I discuss here is autonomy. I bypass the question of whether agents are in a medically diagnosed condition to focus on the normative weight and the implications that autonomy can have. Further, my argument does not engage the question of whether MBME provides medical benefits to its subject (Bublitz 2018: 296). This is also why I use the expression 'enhancement' and not 'treatment.'

3. *Why MBME?*

Of central importance to my paper are the reasons for MBME explained by Matravers and McMahan. First, there is the fact that offenders, due to their

crimes, are liable to treatment that cannot be applied to the general population (Matravers 2018; McMahan 2018: 117-118). In the usual social practice, this is, for example, imprisonment. However, Matravers and McMahan explain in their articles a specific reason for MBME. They take into serious consideration the fact that incarceration is expensive. This introduces a possible justification of MBME in the context of resource allocation between competing rights and values. MBME could be justified if it happens to be less expensive than imprisonment (McMahan 2018: 121-123). Namely, the offender is responsible for his imprisonment. It would be unfair to be frugal about, say, resources for life-saving treatments, in order to offer a dangerous criminal offender who refuses rehabilitation the possibility to choose the way that society will protect people from the violent and harmful acts that he would repeat (McMahan 2018: 121). The argument, as I employ it, raises an allocative issue. The intention is to find the most reasonable balance of the normative strength of competing rights and values.

An assumption in the discussion is that MBE costs will be lower than the costs of incarceration. This is not an unreasonable assumption, given the high costs of incarceration. In the UK, for instance, keeping a person in jail for one year costs the taxpayers a staggering amount of £40,000 (Matravers 2018: 73). For the 86,000 prisoners in England and Wales in 2016, the overall cost was £3.4 billion. We should also note that, while they are in prison, criminal offenders do not earn wages or pay taxes, which represents additional financial burdens for society. Most strikingly, about one-half of the imprisoned criminal offenders re-offend within a year after release and are imprisoned again, bringing about additional costs (between £9.5 and £13 billion) (Matravers 2018: 73-74).

In fact, when commenting on the argument related to the incarceration costs, opponents of MBME admit the assumption that such costs are more significant than MBME costs (Bublitz 2018: 291, 315; Shaw 2015: 1389). Their line of argument is that, even if this were the case, MBME would still not be justified. One possible reason is that the right to oppose MBME is a negative right and, as such, does not imply allocative issues because persons who appeal to it do not require allocation of resources but, merely, non-interventions. The other possible reason is that strong values and rights that are lexically superior in comparison to possible reasons for MBME are undermined by it. Thus, they represent reasons to rebut such practice. These are the arguments that I address in this paper.

Obviously, in order to establish whether MBME could be justified overall in practice through the economic argument, we would require data about the costs of MBME, like, for example, BME-related research and its possible applications. At the present moment, MBME technologies are mainly prospective, so such data remains unavailable. We can only discuss whether costs matter in

principle for justifying MBME, which is, in general, one of the contributions that a philosopher can offer to the debate. The discussion makes sense in virtue of the high costs of imprisonment. This is a problem that, together with the argumentative dialectic accepted by opponents of MBME, as discussed in this paper, justifies looking for alternatives provided they are morally legitimate.

4. *The autonomy objection*

There are at least two versions of the autonomy objection. One of them is based on perfectionism, the other one on respect of basic rights and liberties.

The standard definition of perfectionism establishes that (i) objective goods (values, virtues) exist and (ii) that the state legitimately enforces or favors them. Some perfectionists claim that autonomy is one of them (Wall: 1998). In the MBME debate, the perfectionist objection has appeared in two shapes. One establishes a constitutive link between virtue and autonomy. According to this view, an agent cannot be virtuous if she is not autonomous because autonomy is part of what constitutes virtue.

John Harris has offered early criticisms to BME that can be interpreted as an expression of this perfectionist challenge. Here, like in some other parts of the paper, I consider arguments from the general debate on BME pertinent to the specific question of MBME of criminal offenders. In Harris's view, we must not harm the capacity to exercise choices. This also includes the capacity to choose the bad instead of good, because: "without that freedom, there is no virtue in right action and no evil in wrongdoing. [MBME] would attack agency itself not just prevent bad decisions" (Harris 2016: 98-99). If a person behaves laudably simply because she is programmed not to do otherwise, then she is impeded from attaining virtue (Harris 2011; 2016)[3]. Harris' thesis does not oppose only a complete loss of freedom but also losing a significant portion of it (Harris 2016: 78).

This kind of view is described (but not endorsed) by Michael Hauskeller (2017). In his description of the central perfectionist idea opposed to MBME, a person who does good as a result of MBME is no better than a person who deliberately acts badly. In fact, the latter is preferable.

In the other version, the perfectionist objection points out that virtuous action, because of its complexity and context-sensitivity, requires a subtle capacity of choice and the abilities, with varying moral valence, to realize choices, or, in other words, autonomy. Autonomy is, thus, a complex capacity

[3] To be sure, Harris does not discuss BM of criminal offenders specifically, and I leave aside the interpretative question of what he would say about that particular case. It is nonetheless possible to extend his arguments to the question of MBME of criminal offenders.

needed to have the possibility to manifest virtuous behavior.

John Harris (2016) and Katrina Sifferd, in her virtue ethics proposal (2016), explain why the removal of traits, dispositions, and capacities of criminal offenders reduces them to a less developed stage[4] by limiting the range of options they can choose from. In other words, there are cases when precisely manifestations of traits, dispositions and capacities that, according to my description, are candidates to be removed from persistent offenders are contextually virtuous**.** For example, aggressivity is usually condemnable, but it can be required, when needed to protect victims of maltreatment. Moral enhancement is not achieved by removing traits, dispositions, and capacities that have, in general, negative valence, and therefore, by eliminating some options for the criminal offender. This is so because their valence changes contextually. Thus, moral enhancement is realized by developing the capacity to exercise proper choices, i.e., to choose and practice behaviors rather than limiting the range of options. In other words, moral progress requires autonomy and stands damaged by simplifying the range of possible choices.

Harris's and Sifferd's theses could be related to a generalization of Jacob's arguments (2016) and further supported by it. He speaks about the punishment of criminal offenders and focuses on long-term incarceration. Additionally, his concern is about traits linked to agential capacities in civil society, related to the functioning of a liberal order. But there are no reasons not to extend the discussion to other forms of sanction, like MBME. The central thesis is that punishment must be careful about not harming the character of convicted criminal offenders in a way that disables them to be self-determining agents and valid participants in civil society. The argument is linked to Harris's and Sifferd's theses because they say that alleged MBME deprives of virtuous social interaction capabilities and, thus, presumably, of capabilities needed for virtuous participation in civil society. Therefore, MBME could not be accepted in virtue of Jacob's argument.

The present perfectionist objection might be irrelevant for the conception and goal of MBE that I discuss in the present paper. In some sense, the goal of MBE that is in my focus is comparatively modest. As I present it through the God Machine example, it consists of impeding agents to misbehave. On the other hand, the perfectionist objection that we read in Harris's and Hauskeller's texts is that MBE impedes us to live well.

However, it is visible, for example, through Harris's refusal of God's Machine (Harris 2011) that the perfectionist objection targets even the kind of BME that I have in mind. The objection's point is that the kind of MBE that I have in

[4] I interpret perfectionism and virtue theory as correlated. Perfectionism is the political side of virtue theory, in the present context.

mind damages the perfectionist moral goal, which affirms human virtues and living a good human life. This goal, in the perfectionist objection to MBME, is rendered impossible by it. By trying to ensure a decent life, we impede virtue, and we impair capacities for virtuous participation in social cooperation

The other kind of objection to MBME appeals to human rights and liberties. The thesis is that autonomy is significant and deserves to be protected through a right that has strong lexical priority. This is true for personal autonomy (Bublitz 2018) and autonomy as non-domination (Hauskeller 2017).

Protection of autonomy as non-domination has its rationale in safeguarding an agent's capacity to be the author of her life. The intention is to protect moral capacities and the rationality of a person from possible abuses in a social and political context.

The right to the integrity of what I denote by 'personal autonomy' is constituted or backed up by several familiar rights in liberal democratic constitutional orders (Bublitz 2018: 300-302). Recognizing the right to make rational and free choices and the right of a person to determine the norms that lead her behaviors is a distinctive feature of such orders. In general, the right to personal autonomy is supported by the idea that the state's legitimacy to interfere with persons' freedom lies in its role to regulate conflict. But such disputes involve external behavior. They can be legitimately controlled through external coercion and not by modifying agent's internal capacities (Bublitz 2018: 306; Jacobs 2016).

Further, we see the relevance of autonomy, as it is defined by Bullock (2018: 162, 165-166) and Harris (2016: 78), as the basis of rights and liberties with strong lexical priority in liberal conceptions of justice, if we consider its relevance for the two moral powers described by Rawls. This is visible, in particular, for the capacity for a conception of the good. This is "the capacity to form, to revise, and to rationally pursue a conception of one's rational advantage, or good" (Rawls 2005: 19). It implies elements present in the definition of autonomy. Precisely, the capacity for a conception of the good implies, for example, the capacity to choose actions through rational and free choices (Harris 2016: 78), as well as "the ability to determine for oneself […] considerations and principles on which to act" (Bullock 2018: 162). But, then, we see why a reasonable and rational person can endorse a strong right protective of autonomy. Like Rawls says, people have a higher-order interest in developing and exercising the two moral powers (Rawls 2005: 106). Because of the fact that autonomy as defined above is implied by at least one of the two powers, we have a basis for determining it as protected through a strong right. Of course, in the case of criminal offenders that refuse rehabilitation, society can legitimately intervene to impede behaviors that they choose. Again, the target is represented by their behavior and not by their internal capacities, as the capacity to establish for themselves "considerations and principles on which to act".

5. *Lexical order or a flexible model?*

I move now to the argument that considers the costs of imprisonment as a reason for MBME. It might be objected that the debate on MBME of criminal offenders should not be set up in an allocative framework. The challenge can appear in two forms. First, the objection would be that there is no allocative question since convicted criminal offenders do not demand any resources to protect their autonomy. They only require a negative right that is satisfied by mere abstinence from imposing MBME. The objection says that contrary to positive rights, such rights are not disputable through economic arguments.

This kind of argument has already been criticized by Stephen Holmes and Cass Sunstein (1999) and Colin Farrelly (2007). Their theses are that, based on the facts of real-life, protection of negative rights costs. In a condition of limited resources, we must consider their protection through a proper balance of costs relative to the protection of costs of other rights. "Putting a price tag on such central guarantees" (Bublitz 2018: 315) is thus, a reasonable consideration conformed to a necessity of real life. It "runs counter to the idea of human rights" (Bublitz 2018: 315) only if we do not take into consideration real-world constraints. We see the relevance of costs in the specific case of MBME, as well. We must not lose sight of the fact that a criminal offender who refuses rehabilitation represents a persistent threat to other people. He is legitimately impeded for being a menace to society (not impeding him is not a legitimate option, in consideration of other people's rights). The usual way to do this is incarceration. But imagine that an alternative – MBME – is available and less costly. The prisoner who refuses BME, thus, opts for incarceration. Such an option would cause avoidable economic burdens to society for the sake of maintaining the criminal offenders' traits, dispositions, and capacities to act. In circumstances of limited resources, the orientation of resources for satisfying the convicted criminal offender's goal also implies the denial of resources to other agents for their needs. Consequently, we have a social issue in the context of resource allocation for the protection of rights.

Supporters of liberty rights can reply by affirming their lexical superiority. They could say that, of course, such rights imply costs. But their normative strength is such that they need absolute protection, nonetheless (Bublitz 2018: 315). In virtue of this argument, other rights could only be protected after these rights have been fully protected. This argument is familiar in political philosophy, and it corresponds to Rawls' lexical order. According to this thesis, to put it simply, social and economic justice questions come to the agenda only after questions of basic liberties have been fully managed (Rawls 1999: 37-38, 53-54).

The question that appears is whether adopting a rigid principle of lexical

order is the most reasonable alternative. Or, is it more reasonable to opt for a more flexible model that admits trade-offs? In the present context, for some, the strength of autonomy is such that it defeats other rights and, in effect, the financial rationale that would support MBME were it less costly than incarceration. Others deny such strength to autonomy. The question must be assessed through a proper model of public justification.

6. *Public reason*

At this point, I introduce a distinctive and pivotal element in the debate so far not employed by other contenders in the dispute. I suggest applying to the allocative formulation of the MBME of dangerous criminal offenders John Rawls's model of public justification, the theory of public reason (Rawls 2005). This model's core idea is that public decisions are justified if, and only if, they are justified through reasons that all reasonable persons can accept as free and equal. Included in such reasons are some political ideas, like the idea of society as a fair system of cooperation among free and equal citizens on the basis of reciprocity, some basic rights, liberties, and opportunities, their priority, the means to make effective use of them, and truths and methods of science when these are not controversial. These reasons are shared among reasonable citizens as free and equal, and in the process of public justification, they are referred to as valid public reasons (Rawls 2005: 212-254).

The alternative approach is to identify a favorite moral conception or theory of justice. From this, we derive prescriptions for disputed issues despite whether reasonable persons disagree with it and cannot accept the policy's justification. I assume that this is a wrong approach. There are various explanations for rejecting this approach and the requirement to employ public reason in public justification (Quong 2014: 270-275). One of these is that by enforcing a decision based on reasonably contested reasons, one person, a group, or even the majority of people would take the position of authority concerning other agents as interpreters of the truth (Gaus 2011; Ferretti 2018; Rawls 2005). Other agents would be treated as less than equal, and the basis that earns them the status of free and equal citizens - namely, their moral and epistemic capacities and their responsible use of them - would be pushed to the side. This holds even if the doctrine were true since there would still be reasonable pluralism about such truth. I find this an important consideration. I indicate an additional one, represented by Quong's theory (Quong 2011; 2014: 273-275). The rationale for public reason consists in the support it gives society as a fair system of cooperation among free and equal citizens. This is represented by the requirement to justify, at the very least, public decisions that concern basic rights, liberties, and

opportunities through the kind of reasons listed above.

Notably, public reason does not need to aim to achieve a unique reasonable answer in each dispute "based on decisive public reasons that defeat all competing considerations, and which each citizen is expected to endorse" (Williams 2000: 209). The method of public justification proposed by Rawls can appeal to the resolution of such cases, where public decisions remain undetermined (Williams 2000: 209). We can still hope that further reflection and research will help us to overcome the situation. But, sometimes, rational underdetermination remains persistent. In such cases, public reason obtains relevant achievements, nonetheless. First, although we do not have decisive reasons, i.e., reasons that justify a unique answer to each reasonable person, we have at our disposal undefeated reasons. Such are reasons that a reasonable person can endorse. Second, although we cannot establish, through undefeated reasons, uniquely required decisions, we have arrived at a set of eligible decisions (Williams 2000: 209). The achievement is significant because we have, at least, warranted that the final decision will not be unreasonable.

I illustrate this with cases in a pandemic. We need to make a public decision. We have proposal P1 based on pseudoscience and conspiracy theories. We have two further alternative proposals, P2 and P3, based on undefeated valid scientific reasons. Because both P2 and P3 are justified through undefeated but not decisive, valid public reasons, we do not have a unique response that each reasonable person must accept. We have, nonetheless, done significant work in distinguishing between reasonable proposals P2 and P3 on one side, and unreasonable P1, on the other side. P2 and P3 are eligible proposals, while P1 is not eligible.

But, in some situations, there might still be pressure to make a public decision. In such a case, we can recur to a fair resolution, like a fair democratic procedure (Williams 2000: 210). The decision is fair, in virtue of the procedure, and reasonable, at the same time, because it is represented by choice between reasonable eligible proposals.

Coming back to the present paper's topic, we can assume that reasonable agents accept one of the reasons employed in the MBME debate, autonomy, as normatively stringent. For them, autonomy is a valid public reason for justifying public decisions. However, for some of them, autonomy implies a right that always has lexical superiority and, thus, deserves absolute protection. For others, supremacy can be negotiated to some extent. For them, the right to health care, for example, can in some cases be normatively stronger. We have, thus, opposite proposals, both of them eligible, because they are both supported by reasons that all reasonable persons can accept as free and equal. In such circumstances, legitimacy belongs to the decision that is victorious among eligible proposals in a fair procedure.

One could question the use of the public reason test in the present context. For instance, one could note that Rawls himself has never written about criminal offense and punishment. Further, a famous criticism of his theory affirms that the exclusion of some topics from his opus is not accidental. Instead, the argument says that it testifies the limits of his paradigm (Nussbaum 2006). On the contrary, I think it is possible and valuable to extend Rawls's doctrine to topics not embraced in his opus. I do this for the theory of public reason and for a topic that enters the field of criminal justice, side-stepped by him.

Rawls indicates the proper application of public reason to "cases involving the constitutional essentials, and, also, in other cases, insofar as they border on those essentials and become politically divisive" (Rawls 2001: 117). For example, abortion is a case when public reason needs to be applied, because, although it is not strictly a constitutional essential, it borders on it and is politically divisive. In my view, MBME can be such a case. It is related to questions of basic rights and liberties because it strongly enters into the domain of liberty of conscience and the "liberties specified by the liberty and integrity (physical and psychological) of the person" (Rawls 2001: 44). At the moment, we cannot say that it is politically divisive, like abortion, in public at large. But the reason is that, for the moment, it is mainly a conceivable possibility. However, the actual academic dispute shows that it can become politically divisive if the relevant biotechnologies will have a clear perspective of being available.

A further challenge could be represented by the fact that the method of public reason is proper to mutual justification among reasonable persons. Such are persons who, among else, recognize other persons as free and equal and who are disposed to establish with others fair cooperation on terms of reciprocity (Rawls 2005). The question is, why are we, and how could we be, obliged to justify to clearly unreasonable persons, like Hannibal and Anders, public decisions on public reason terms.

To answer this objection, it is important to establish the subjects to which public justification is addressed. Notably, we are not obliged to justify principles and public decisions to unreasonable persons. In other words, we are not obliged to address them justification based on reasons that they can accept. Reasonable persons establish principles of justice and make related public decisions for them. But, reasonable persons must not do this arbitrarily. Instead, they must properly justify reasonable principles and public decisions. This is required to avoid the challenge addressed by Martha Nussbaum to Rawls's political philosophy (Nussbaum 2006) that claims how Rawls neglects an entire set of relevant categories of subjects. For instance, the mentally impaired or disabled people that, according to our considered judgments, deserve justice considerations. One could say the same for criminal offenders.

But how can reasonable persons extend justification of principles and public decisions to subjects who deserve consideration of justice but are not qualified to take part in the justificatory process? In my view, and in coherence with Rawls's original view of public justification, reasonable persons who participate in this justificatory process must apply the method of public reason. Similarly to how they do when they justify principles and public decisions that apply to themselves. In other words, they must justify to each other how the principles and public decisions stand applied on non-reasonable persons by relying on reasons that each reasonable person can accept. In this process, they must use the same reasons, or reasons coherent with these reasons, that they usually employ when principles and public decisions that apply to them are concerned. I illustrate this through an example related to the present paper.

Let's assume that a, at least pro tanto, valid public reason for the justification of public decisions applied to reasonable persons is represented by the following principle. States can legitimately coercively regulate only the behavior of competent persons but not their inner states. A sufficient condition for being a competent person is that the person can choose her actions through rational and free choices, determine for themselves evaluative reasons, etc. At least pro tanto, the consequence is that we must not regulate competent persons' inner states through coercive means even though they are not reasonable.

I show, now, how public reason functions in practice. I turn to the question of whether the thesis that autonomy is so normatively strong that it trumps competing reasons in resource allocations is publicly justified and, as a consequence, whether MBME of criminal offenders is defeated. In what follows, I distinguish between two different debates on the normative weight of autonomy and the allocative question regarding MBME: perfectionist arguments and human rights and liberties arguments.

7. *Perfectionism and autonomy*

I now discuss the perfectionist debate on autonomy's normative weight as a defeater of MBME of criminal offenders. Perfectionist conceptions that place autonomy at the core of their accounts represent a possibly decisive defeater of the economic argument for MBME of criminal offenders. If autonomy has such immense normative weight, it would have lexical priority. In other words, the allocative dispute with other rights and values that can be appealed to by supporters of MBME would never arise.

The classic public reason reply to such a thesis is that perfectionism cannot function in public justification because it represents a controversial philosophical doctrine not shared by citizens who reason as free and equal (Quong 2011).

There are, however, forms of perfectionism that are not evidently liable to such an objection. Collis Tahzib has proposed a kind of perfectionism, called political, by analogy to Rawls's political liberalism (Tahzib 2022). It preserves the Rawlsian public reason constraint that fundamental political decisions must be justified through reasons that all reasonable citizens can accept. But contrary to the typical political liberal view (Quong 2011), Tahzib affirms that perfectionist values can satisfy this condition. He assumes that arts and sciences are among perfectionist values that reasonable persons can share. Can we assume that autonomy is among these values, as well? In my view, we cannot assume that all reasonable persons will attribute value to autonomy in the perfectionist sense, at least not in the cases relevant for MBME, and, thus, we cannot employ it as a valid public reason. But before concluding this, I must analyze the thought of authors who do not share this view.

Harris (2016) and Sifferd (2016) offer reasons to endorse autonomy as a virtue. They explain that virtuous action, because of its complexity and context-sensitivity, requires a subtle capacity of making and realizing choices, or, in other words, autonomy. This is because the valence of actions and dispositions changes. MBME, thus, implies an enormous loss because it deprives agents of the capacity needed to practice virtuous behavior, or, at least, it strongly limits this capacity.

Still, Harris and Sifferd do not offer conclusive reasons for publicly justifying the needed strength of the value of autonomy in the present case. Their central thesis applied to the present context is that, after some traits, dispositions, and capacities have been removed from persistent criminal offenders, they will lose capacities that are sometimes morally laudable or the capacity to choose to practice these capacities out. They will thus be deprived of the potentialities to be virtuous in specific contexts. However, in my view, such potentialities are ephemeral in the present case because, for example, Anders and Hannibal will never exercise them or will exercise them only in limited cases. Such cases are irrelevant if compared with the horrible acts that they are disposed to perform.

Harris and Sifferd shape the debate so that it seems how the salient choice is between two alternatives—on the one hand, leaving agents the capacity to make and practice fine-grained reasonable choices (that implies autonomy). And on the other hand, making them people with restricted options (which limits their choices, including, on some occasions, to do what morality requires). But, due to agents' resistance, in the present discussion, the salient alternative stands between the option of restricting their choices and leaving them to remain persistent and ferocious criminals. The criticism to MBME cannot be that the capacities removed from criminal offenders resistant to rehabilitation could be helpful in some cases. Therefore, we must leave agents with these capacities and

the ability to choose when to exercise them. In fact, because of such agents' persistent dispositions, decisions, and preferences, such capacities will be mainly used for heinous acts, because such cases are marginal in their lives. Therefore, their virtuous potentialities are ephemeral. The loss of these capacities and the capacity to decide when to practice them are not a loss of virtue in this specific case. Thus, applying MBME does not cause a loss of virtue in this specific case.

The option that remains to perfectionists is to affirm a constitutive link between autonomy and perfection, or virtue. The objection to MBME is that, by utilizing it, we deprive a person of virtue because we remove a feature that represents a necessary component of the good human life. But such an assumption does not satisfy the public reason test. Imposing the idea that "ways of life have value only if they are freely chosen" (Barry 1995: 130) as a basic normative principle represents the imposition of a partisan view. From the public reason perspective, the problem is that some reasonable persons cannot endorse such an idea.

Further, Harris's and Seffird's argument in opposition to MBME can be opposed even by those who endorse the view that autonomy is a necessary component of a fully good human life. Namely, to be an objection to MBME, the thesis must be that autonomy does not only constitute virtue as one of its elements but that its presence implies virtue. It follows that a person who has the capacity to exercise choices and the abilities to practice them but who, persistently, has no, or has rare, intention to behave properly, is, at least in a relevant sense, in a virtuous condition. This is needed to say that being a bad person A is better than being an inoffensive but cooperative member of society, B if B is harmless due to MBME and restriction of autonomy. However, we might reasonably disagree whether restricting the autonomy of those who would mainly choose horrible options constitutes a moral loss.

Steven Wall, for example, offers reasons to think that it does not (Wall 2008). In fact, Wall does not speak in terms of public reasons. He affirms his theory as the simply correct one. But we can assume that his thesis, about how autonomy persistently exercised for the bad is not virtuous, stands to be a valid challenge in public justification. This is a thesis that a reasonable person can endorse. Favoring autonomy as an absolute overriding value despite such disagreement is, thus, not publicly justified through the lens of public reason.

This leads us to a conclusion for the present allocative issue. The perfectionist reasons described above do not provide legitimate public justification to protect the autonomy and mental integrity of criminal offenders who refuse rehabilitation at the expense of other normative demands.

8. *The allocative question*

I now discuss the allocative questions related to MBME by drawing from the discussion on rights and liberties as basic considerations. I start with the position that attributes personal autonomy with the status of superior right in lexical order.

This is a reasonable position. It appeals to a right that can be endorsed by all persons as free and equal. We see this, for example, through Bublitz's explanation shown above, as well as through the role of autonomy, as defined by Bullock (2018: 162, 165-168) and Harris (2016: 78), in Rawls's liberal theory of justice. But even though all reasonable persons can accept the appeal to autonomy as a valid public reason, reasonable persons can refuse its rigid priority in lexical order.

Farrelly (2007) has objected to the endorsement of such a rigid lexical principle. His specific target is represented by the rigid lexical priority that Rawls attributes to basic liberties. He says that the endorsement of such a rigid lexical priority implies that, in real life, important needs can never come to the agenda in virtue of real-life conditions. Instead, he proposes a more flexible view. According to such a view, we must not assume a rigid lexical order but consent trade-offs based on reasonable judgments sensitive to context.

A similar proposal can be applied to the MBME dispute. We could say that we must resolve the question of resource allocation when we address competing claims through context-sensitive, reasonable judgment. Such can be claimed for the protection of autonomy, on one side, or the protection of the right to healthcare, or the right to good education, on the other side. But, reasonable agents can diverge on the normative weight they attribute to autonomy and other rights and values. This is crucial in the present debate on MBME that I frame as an allocative issue. Some can consider autonomy as protected by a right superior in lexical order, that, in possible conflicts, overrides all other rights in public decision-making procedures. But the opposite view is reasonable as well. One can, legitimately from the public reason's viewpoint, attribute stronger weight to competing rights in some of these conflicts. Specifically, while some can think that autonomy is an overriding right in the cases of criminal offenders that refuse rehabilitation, others can think that it could be sacrificed through MBME for the sake of competing rights. First, this is so because the rights that compete with autonomy, such as healthcare or good education, represent reasons we can accept as reasonable persons. Further, one can adduce reasons that reduce the weight of the claim of criminal offenders. For example, one can say that the autonomy of criminal offenders can be sacrificed through MBME for the sake of competing rights in virtue of the poor way it is exercised (i.e., maintaining

criminal dispositions and capacities). Aside from the awful way in which serious criminal offenders exercise their autonomy, there is an additional reason to limit supporting it. It would be a waste of resources to employ them into cultivating dispositions for acts that will be impeded by means of even stronger limitations of freedom, like imprisonment. A further reason that we need to consider here is that the criminal offender is responsible for his condition. In contrast, his opponents in the allocative dispute, as a child in need of life-saving treatments, by assumption, are not (McMahan 2018: 121).

I argued for the conclusion that attributing overriding strength to the right to personal autonomy is reasonably contestable, and, thus, it cannot be considered decisive against the MBME of criminal offenders. However, I do not arrive at a conclusion that adjudicates between the competing theses – those in favor and those opposed to MBME. I merely show that we are within the bounds of reasonable pluralism. Reasonable persons can disagree about such theses. Each side can, sustained by valid public reasons, affirm either the thesis that autonomy must be safeguarded no matter the cost (in which case MBME of criminal offenders can have no legitimacy), as well as that the protection of autonomy must be balanced with other rights in an allocative dispute, and that, sometimes, such rights can deserve priority. The fact that there is no unique reasonable answer is not a problem for the public reason theory of justification, as shown in section 6. There, I have remarked that the aim of public justification is obtained when we warrant that public decisions are reasonable and fair. And in cases like the present one, where we do not have a unique reasonable answer, we use a two-step procedure. First, we arrive at a step of eligible valid public decisions. Each of them is justified through reasons that each reasonable person can accept. Second, we make the final decision among them in a fair procedure.

I turn, finally, to the opposition to MBME that is based on the idea of non-domination as advanced by Hauskeller (2017). As I have described in section 4, autonomy, in this context, is interpreted socially as the condition of not being dominated by others. In defense of the legitimacy of MBME, one can appeal to the interpretation of non-domination coherent with the Rawlsian theory of public justification. Non-domination is achieved when decisions are justified through public reason. This is so, even when the subjects of decisions do not factually consent to them. What matters is idealized consent. This is the consent that each reasonable person would give in virtue of her reasonableness and not factual consent.

Consequently, this is true for MBME, as well. Subjects of MBME are interpreted as co-authoring the decision because it is justified through reasons that they can all accept as reasonable, free and equal persons. Such are, first, rights that compete with autonomy in an allocative question, and that they accept as

reasonable, free and equal, and, second, the verdict of a fair procedure of choice among eligible proposals. Namely, when valid public reasons do not provide a unique reasonable answer, legitimacy derives from a fair procedural decision that is the expression of equality of all persons.

9. *Conclusion*

In conclusion, I have introduced some novelties into the debate about the MBME of criminal offenders. The first is methodological. In the current state of the MBME debate, authors typically proceed from their favorite moral views, broadly conceived, or from their favorite insights on values or rights. From these, they derive recommendations that they consider publicly justified. Instead, I employ the Rawlsian model of public reason that employs only reasons that all persons can accept as free, equal, and epistemically responsible. This approach pays attention to the pluralism among persons. It respects their freedom and equality by refraining from recommending proposals that are justified through reasons that some citizens cannot accept as reasonable, free and equal.

Second, I have offered further support to the thesis that costs matter when deciding about the MBME of criminal offenders. I, however, only endorse the principled version of this claim since we cannot know the costs of practices that are mainly prospective. But the discussion about the principle serves as an indication that moral relevance should be attributed to costs. Such an indication can be relevant to institutions that could be involved in decisions on whether to support the research of such practices and under what conditions.

Third, I have demonstrated that the perfectionist appeal to autonomy is not an absolute defeater for the MBME of criminal offenders. Such appeal is not admissible in public justification, from the Rawlsian public reason perspective, because it is linked to conceptions of good and value that are not shared by all reasonable agents.

Fourth, I have demonstrated that the non-perfectionist appeal to autonomy is admissible in the human rights and liberties context because it refers to a right that is part of the political culture of a society of free and equal citizens. But there can be reasonable disagreement among citizens as free and equal about its relative normative strength when we need to balance it with other rights in allocative disputes. Thus, we lack a conclusively victorious decision. In reasonable pluralism circumstances, the best that we can achieve is for final legitimacy to be derived from a fair decision-making procedure among eligible proposals.

As far as the arguments in the present paper are concerned, MBME of criminal offenders is neither defeated nor conclusively justified. The decision

pertains to a fair deliberation in the decisional process, like with other allocative questions, as indicated, for example, by Daniels and Sabin in healthcare justice (Daniels *et al.* 2008)[5].

Elvio Baccarini
(University of Rijeka)
ebaccarini@ffri.hr

References

Barry, Brian, 1995, *Justice as Impartiality*, Oxford University Press, Oxford.

Bublitz, Jan Christoph, 2016, "Moral Enhancement and Mental Freedom" in *Journal of Applied Philosophy* 33: 88-106.

—, 2018, "'The soul is the prison of the body': Mandatory Moral Enhancement, punishment, and rights against neurorehabilitation" in Birks *et al.*, eds., *Treatment for crime: Philosophical essays on neurointervention in criminal justice*, Oxford University Press, Oxford: 289-320.

—, *et al.*, 2014. "Crimes against minds: On mental manipulations, harms and a human right to mental self-determination" in *Criminal Law and Philosophy* 8: 51-77.

Bullock, Emma, 2018, "Moral paternalism and neurointerventions" in Birks *et al.*, eds., *Treatment for crime: Philosophical essays on neurointervention in criminal justice*, Oxford University Press, Oxford: 159-177.

Chew, C. *et al.*, 2018. "Biological interventions for crime prevention" in Birks *et al.*, eds., *Treatment for crime: Philosophical essays on neurointervention in criminal justice*, Oxford University Press, Oxford: 11-43.

Danaher, John, 2018, "Moral enhancement and moral freedom: A critique of the little Alex problem" in *Royal Institute of Philosophy Supplement* 83: 233-250.

Daniels, Norman *et al.*, 2008, *Setting limits fairly: Learning to share resources for health*, Oxford University Press, Oxford.

Douglas, Thomas, 2014, "Criminal rehabilitation through medical intervention: Moral liability and the right to bodily integrity" in *The Journal of Ethics* 18: 101-122.

[5] I have contracted many debts with colleagues and friends that helped me extensively while I was working on drafts of the present paper. I would like to very much thank the audiences of the talks organized by Carla Bagnoli at the University of Modena (in particular Massimo Reichlin and Sarah Songhorian, as well as Carla Bagnoli herself for the comments at the presentation, and for subsequent written comments), by Thomas Douglas at the University of Oxford and by Sergio Filippo Magni at the University of Pavia. I thank Sara Amighetti, Richard Arneson and Julian Savulescu for their helpful comments when I presented the paper at the summer school in Rijeka, Viktor Ivankovic and Collis Tahzib for their written comments and an anonymous reviewer of an earlier version of the paper. As usual, I received great help from my friends at the Department of Philosophy in Rijeka: Ivan Cerovac, Tomislav Furlanis, Ana Gavran Milos, Iva Martinic, Marko Jurjako, Kristina Lekic Baruncic, Luca Malatesti, Aleksandar Susnjar and Nebojsa Zelic

—, 2018. "Neural and environmental modulation of motivation: What's the moral difference?" in Birks *et al*., eds., *Treatment for crime: Philosophical essays on neurointervention in criminal justice*, Oxford University Press, Oxford: 208-223.

Farrelly, Colin, 2007, *Justice, Democracy and Reasonable Agreement*, Palgrave MacMillan, Basingstoke.

Ferretti, Maria Paola, 2018. *The Public Perspective*, Rowman & Littlefield, London.

Giubilini, A. *et. al*., 2018, "The artificial moral advisor: The ideal observer meets artificial intelligence" in *Philosophy and Technology* 31: 169-188.

Harris, John, 2011, "Moral enhancement and freedom" in *Bioethics* 25: 102-111.

—, 2016, *How to be Good: The Possibility of Moral Enhancement*, Oxford University Press, Oxford.

Hauskeller, Michael, 2017, "Is it desirable to be Able to do the undesirable? Moral bioenhancement and the little Alex problem" in *Cambridge Quarterly of Healthcare Ethics* 26: 365-376.

Holmes, S., *et al*., 1999, *The cost of rights: Why liberty depends on taxes*. New York: W.W. Norton.

Jacobs, Jonathan, "Character, Punishment, and the Liberal Order" in Masala *et al*., eds., *From personality to virtue: Essays on the philosophy of character*, Oxford University Press, Oxford: 9-34.

Liberto, Hallie, 2018, "Chemical castration and the violation of sexual rights" in Birks *et al*., eds., *Treatment for crime: Philosophical essays on neurointervention in criminal justice*, Oxford University Press, Oxford: 196-207.

Lippert-Rasmussen, Kasper, 2018, "The self-ownership trilemma. Extended minds, and neurointerventions". In *Treatment for crime: Philosophical essays on neurointervention in criminal justice*, Birks D. & Douglas, T. eds., 140-158. Oxford: Oxford University Press.

Matravers, Matt, 2018. "The importance of context in thinking about crime-preventing neurointerventions" in Birks *et al*., ed., *Treatment for crime: Philosophical essays on neurointervention in criminal justice*, Oxford: Oxford University Press: 71-93.

McMahan, Jeff, 2018, "Moral liability to 'crime-preventing neurointerventions" in Birks *et al*., ed., in *Treatment for crime: Philosophical essays on neurointervention in criminal justice*, Oxford University Press, Oxford: 117-123.

Nussbaum, Martha, 2006, *Frontiers of Justice: Disability, Nationality, Species Membership*, Harvard University Press, Cambridge, Mass.

Quong, Jonathan, 2011, *Liberalism without Perfection*, Oxford University Press, Oxford.

—, 2014, "On the idea of public reason" in Mandle *et al*., eds., *A Companion to Rawls*, Wiley Blackwell, Oxford: 265-80.

Persson, Igmar, *et al*., 2012, *Unfit for the Future: The Need for Moral Enhancement*, Oxford University Press, Oxford.

Rawls, John, 1999, *A Theory of Justice*, Harvard University Press, Cambridge, Mass.

—, 2000, *Lectures on the History of Moral Philosophy*, Harvard University Press, Cambridge, Mass.

—, 2001, *Justice as Fairness: A Restatement*, Harvard University Press, Cambridge, Mass.

—, 2005, *Political Liberalism*, Columbia University Press, New York.

Ryberg, Jesper, 2012, "Punishment, pharmacological treatment, and early release" in *International Journal of Applied Philosophy* 26: 231-244.

Savulescu, Julian, *et al.,* 2015, "Moral enhancement and artificial intelligence: Moral AI?" in Romportl *et. al*., eds., *Beyond artificial intelligence: The disappearing human-machine divide*, Kluwer, Dordrecth: 79-95.

—, *et al*., 2012, "Moral enhancement, freedom, and the god machine" in *The Monist* 95: 399-421.

Shaw, Elizabeth, 2011, "Cognitive enhancement and criminal behaviour" in van den Berg *et al*., eds., *Technologies on the stand: Legal and ethical questions in neuroscience and robotics*, Wolf Legal Publishers, Nijmegen: 187-204.

—, 2015, "The use of brain interventions in offender rehabilitation programs: Should it be mandatory, voluntary, or prohibited?" in Clausen *et al*., eds., *Handbook of Neuroethics*, Springer, Dordrecht: 1381-1398.

—, 2018, "Against mandatory use of neurointerventions" in Birks *et al*., eds., *Treatment for crime: Philosophical essays on neurointervention in criminal justice*, Oxford University Press, Oxford: 321-337.

—, 2019, "The right to bodily integrity and the rehabilitation of offenders through medical interventions: A reply to Thomas Douglas" in *Neuroethics* 12: 97-106.

Sifferd, Katrina, 2016, "Virtue ethics and criminal punishment" in Masala *et al*., eds., *From personality to virtue: Essays on the philosophy of character*, Oxford University Press, Oxford: 35-61.

Tahzib, Collis, 2022, *A Perfectionist Theory of Justice*, Oxford University Press, Oxford.

Wall, Steven, 1998, *Liberalism, Perfectionism and Restraint*, Cambridge University Press, Cambridge.

Williams, Andrew, 2000, "The Alleged Incompleteness of Public Reason" in *Res Publica* 6: 199-211.

William MacAskill

What We Owe the Future: A Million-Year View

London: Oneworld Publications, 2022; hardback, 352 pp., £20.00,
ISBN: 9780861546138

B.V.E. Hyde

Moral circle expansion has been occurring faster than ever before in the last forty years, with moral agency fully extended to all humans regardless of their ethnicity, and regardless of their geographical location, as well as to animals, plants, ecosystems and even artificial intelligence. This process has made even more headway in recent years with the establishment of moral obligations towards future generations. Responsible for this development is the moral theory – and its associated movement – of longtermism, the bible of which is *What We Owe the Future* (London: Oneworld, 2022) by William MacAskill, whose book *Doing Good Better* (London: Guardian Faber, 2015) set the cornerstone of the effective altruist movement of which longtermism forms a part.

Longtermism was perhaps first brought to prominence by Toby Ord in *The Precipice* (London: Bloomsbury, 2020) who defined it as a 'moral re-orientation toward the vast future' (p. 52). Longtermists argue that the (utilitarian) principle of impartiality, or the equal consideration of interests, means that, as Peter Singer, perhaps the principal utilitarian philosopher of our time, says: 'it makes no moral difference whether the person I can help is a neighbor's child ten yards away from me or a Bengali whose name I shall never know, ten thousand miles away' (*Philos. Public. Aff.* vol. 1, no. 3, pp. 229-243; 1972). For Mr. MacAskill, 'distance in time is like distance in space' (p. 10) so, if we are to care about a Bengali ten thousand miles away, then we ought to care about one ten thousand years into the future.

There are some problems with the utilitarian principle of impartiality – and they are not new problems either – none of which are mentioned by Mr. MacAskill, but he seems to be aware of them, because he clunkily adds to his justification of longtermism a deontological principle completely opposed to utilitarianism. He says of future generations that, 'if we recognize they are real people… then we have a duty to consider how we might impact the world they inhibit' (p. 19). This is a rehashed version of Immanuel Kant's 'formula of humanity' which he laid out in the *Groundwork of the Metaphysics of Morals* (Riga:

ISSN (print) 2281-8618-ETS

Johann Friedrich Hartknoch, 1785): 'act that you treat humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means' (p. 429). It does not seem to strike Mr. MacAskill as problematic that Immanuel Kant was referring to conscious persons with moral autonomy who are, crucially, alive, and not to the mere idea of possible people who do not exist but might yet still.

Mr. MacAskill thinks that there is a 'tyranny of the present over the future' that needs to be toppled (p. 9). However, one of the chief difficulties for longtermism is that future people do not exist yet, so he must justify why it is good to make happy people. To do so, he tackles the 'intuition of neutrality' (p. 171) which is, in the words of Jan Narveson (*Monist*, vol. 57, no. 1, pp. 62-86; 1973), that 'we are in favour of making people happy, but neutral about making happy people' (p. 80). Mr. MacAskill has four arguments against this intuition.

The first argument begins with the assumption that our intuition is asymmetrical, meaning that we are indifferent about creating happy people but believe it is morally wrong to bring a miserable new person into existence. If our intuitions truly exhibit this asymmetrical nature, then any argument supporting the notion that it is wrong to create an unhappy person should also apply to the idea that it is good to bring a happy person into the world (p. 172).

The second argument is simply that, because it is intuitive to him that the future is better because of the existence of his happy nephews and nieces, it follows that the world is in fact better with the creation of happy people (p. 172).

The third argument departs from the previous two by relying on empirical findings instead of logical reasoning. He refers to a recent study in psychology that discovered that our moral intuitions regarding the creation of happy or unhappy individuals are actually symmetrical, suggesting that we generally believe it is positive to bring happy people into existence and negative to bring unhappy ones (*Cognition*, vol. 218, art. 104941; 2022).

The fourth argument is that, because a minor shift in timing could have led to a different individual being born instead of you, the sperm responsible for your existence having only a one in two hundred million chance of fertilizing an egg, we are 'like clumsy gods' (p. 174), dramatically altering history's trajectory with each passing moment. From what he calls the 'fragility of identity; (p. 173), the implication is that today's policies will impact the future, not by enhancing the lives of people who would have existed regardless, but by *creating* a new future with individuals who are somewhat happier. Moreover, because it is intuitive that we have indeed improved the future, it must be true that adding people with happier lives is good, thereby disproving the intuition of neutrality (p. 176).

The fifth and final argument offered by Mr. MacAskill is the most sophisticated, but it is not his anyway: he admits by way of an endnote that he takes it

from John Broome's book, *Weighing Lives* (Oxford: Oxford University Press, 2004). Say that in world$_1$ you are not born, in world$_2$ you live in suffering and in world$_3$ you live blissfully. The intuition of neutrality says that world$_1$ is neither better nor worse than world$_2$, which means that world$_1$ is equal in value to world$_2$. What licenses this inference is that John Broome assumes the comparative value relation is complete (§10.1), which means that if something is neither better nor worse than something else, the two are equally as good as one another. From the intuition of neutrality it also follows that world$_1$ is just as good as world$_3$. If values are transitive, which Mr. MacAskill assumes they are, then it follows that world$_2$ is just as good as world$_3$ which, according to Mr. MacAskill, is a 'contradiction' because it cannot be the case that creating a life of suffering is just as good as creating a life of bliss (p. 177). Therefore, it must be good to create good lives.

None of these arguments are sound. More than one begs the question. The strongest is the evidential one, but it does not follow from evidence that we *do* think it good to create happy people that we *should* think so. This runs afoul of David Hume's law, which he explicated in *A Treatise of Human Nature* (London: John Noon, 1739): that no moral statement can be inferred from non-moral ones (bk. iii, pt, i, §1).

Because creating good lives is good, Mr. MacAskill recommends that we ought to have children (p. 187) and to ensure that civilization lasts as long as possible and is as big as possible (p. 188). The bigger the future, the better the future, which is why 'the early extinction of the human race would be a truly enormous tragedy' (p. 189). This is why Mr. MacAskill argues that we are morally obliged to mitigate existential risks, which Nick Bostrom defines as a threat to the premature extinction of intelligent life on earth or the permanent and drastic destruction of its potential for desirable future development (*Global Policy*, vol. 4, no. 1, pp. 15-31; 2013).

The principal existential risks are, according to Mr. MacAskill, engineered pathogens (p. 107), war between great powers (p. 114), climate change (p. 134) and fossil fuel depletion (p. 138). Many futurological researchers, like Mr. Bostrom, in *Superintelligence* (Oxford: Oxford University Press, 2014), are most concerned by existential risk from artificial general intelligence, where humans could be replaced as the dominant lifeform on earth were machine brains to surpass human brains and become superintelligent. Some are skeptical of this alarmism, like Michio Kaku who, in *Physics of the Future* (New York: Doubleday, 2011), said that he believed we will find intelligent robots benevolent and friendly. Mr. MacAskill is both an alarmist and an optimist, for he believes that artificial intelligence might wipe out the human race, but that it still represents intelligent life with moral value, so even its destruction of humanity would not

be a crisis so long as the artificial civilization that advances into the future is not morally bankrupt (p. 87).

Despite the threat of annihilation, Mr. MacAskill thinks that we should be optimistic about the future (p. 193), in part because the world is already good. Mr. MacAskill commissioned psychologists to run a survey which found that, although around 10% of the global population have lives below neutral well-being, most people have positive lives (p. 201). Moreover, the world is getting better. Richard A. Easterlin published a very famous study in a chapter in *Nations and Households in Economic Growth* (New York: Academic Press, 1974) in which he showed that people and countries do not get happier as they get richer over time. However, it has since been revealed that the Easterlin Paradox does not exist. More recent work with better data strongly supports the hypothesis that countries get happier as they get richer (*Brook. Pap. Econ. Act.* no. 1, pp. 1-87; 2008). Likewise, contrary to the common belief, originating with the psychologist Philip Brickman and his colleagues (*J. Pers. Soc. Psychol.* vol. 36, no. 8, pp. 917-927; 1978), that lottery winners are unhappy, Andrew Oswald and Rainer Winkelmann have shown in a chapter in *The Economics of Happiness* (Cham: Springer, 2019) that winning the lottery does increase one's happiness. If the world continues to get richer, we can expect the future to be even happier.

The future can only be good if good values permeate it, though. Values, Mr. MacAskill thinks, can persist for extremely long periods of time through 'value lock-in' (p. 78), of which Confucian influences on the Orient today and Christian influences on the modern Occident are exemplary. The permanence of values is determined by an 'early plasticity, later rigidity' cycle (p. 42). According to Mr. MacAskill, history is like glass that is sometimes hot and sometimes cold. When it is hot, it can be reshaped, but the colder it gets the harder it becomes. As Derek Parfit wrote in his book *On What Matters* (Oxford: Oxford University Press, 2011), we 'live during the hinge of history' (vol. 2, p. 611). The present age is one of plasticity, but longtermists warn that a period of rigidity is on the horizon. What will cause it, Mr. MacAskill says, is artificial intelligence: because it is immortal and has the potential to cause rapid technological progress, whatever values it holds, or whatever values are instilled within it, could last a very long time (p. 83). This means that our values could define the future, which is why changing them for the better is one of the most important longtermist tasks (p. 52).

Really, we should try to avoid value lock-in (p. 88) and have a 'long reflection' (p. 98) where we can work out what a flourishing society would look like. This should give us a 'morally exploratory world' in which better morals win over time such that we converge on the best society (p. 99). There are a few things we need to do to avoid value lock-in. One: we must prioritize the prevention of value lock-in, even at the expense of delaying advancement such as space ex-

ploration or development of artificial intelligence. Two: we must be politically experimental and ensure that our society is culturally and intellectually diverse to avoid premature convergence. Three: we must somehow ensure that cultural evolution results in moral evolution. What we end up with is a 'lock-in paradox' (p. 101): we need to lock-in some institutions and values to prevent a more thoroughgoing lock-in of values.

What We Owe the Future is a well-researched book, bringing to attention lots of diverse and interdisciplinary evidence, interesting facts, and historical cases to support its arguments. It also contains some original empirical research, and several well-designed illustrations have been produced to make some of the more challenging aspects of the book easier to understand and to make some of the more grandiose claims seem even more impactful. The book has its own website (whatweowethefuture.com) where the bibliography is to be found, rather than in the book, which is odd and worth mentioning. The website also contains some supplements, press about the book, and links to established effective altruist organizations that readers are pointed towards in the book, like 80,000 Hours and the Longtermism Fund. Clearly, lots of hard work has gone into the book.

Lots of it, though, is not Mr. MacAskill's. He admits that the book is an extremely collaborative effort and even that 'many sections of the book were essentially coauthored' (p. 247). If you compare his enormous acknowledgements section with the American Psychological Association (APA) author determination guidelines, you might be surprised that only one name is on the front cover of the book. Even the more stringent International Committee of Medical Journal Editors (ICMJE) recommendations would suggest that some of those acknowledged have been cheated out of authorship. Mr. MacAskill is really the book's editor, not the sole author, and there is definitely a looming question over it about the extent to which his claim to sole authorship represents a questionable research practice. You get the impression that the research for the book was done by a team of researchers, whereas the philosophic arguments are the work of the one, which is perhaps why the historical work is much more impressive than the philosophic, which is not well thought out at all.

In fact, Mr. MacAskill's arguments for longtermism represent some of the poorest for what is perhaps the most popular philosophical movement in the world right now. He argues almost entirely by catching the reader in a provocative literary style that has captured so many established academic celebrities like Stephen Fry and Sam Harris. It is the same attractive, optimistic style that was applauded by reviewers such as Amia Srinivasan with respect to some of his earlier books (*Lond. Rev. Books*, vol. 37, no. 18; 2015). But not everyone has been caught in the excitement conjured up by Mr. MacAskill, and some other

reviewers have also criticized his book for being 'replete with highfalutin truisms, cockamamie analogies and complex discussions leading nowhere' (*Wall Str. J.* 26 August 2022) – to which we must add appeals to intuition, inferences from anecdotal evidence, unjustified assumptions, question begging and, of course, the intellectual crime of utter thoughtlessness: more than half the time, Mr. MacAskill is totally unaware of the positions he is committing himself to, and he often prefers cheap tricks in place of proper philosophic argumentation.

Laid plain of his alluring narrative, there is no philosophic substance to the book in the slightest. It is just another episode in the rehashing of an old and outworn utilitarian theory in a contemporary jacket. The ethical wing of effective altruism and longtermism, as they both currently stand, is nothing but utilitarianism with a vocabulary updated to include buzzwords like climate crisis, global poverty, and artificial intelligence. Perhaps these positions on ethics, philanthropy and global priorities can be put right, but it is very unfortunate that a foundational text is so inadequate; in this regard, the movement's future looks bleak, and it will be forced to choose between objectivity and dogma at its current rate. What we owe the future is a better explanation. Or, at least, William MacAskill does.