IfILNOVA

ESSAYS ON VALUES

VOLUME 3

Maria João Mayer Branco
João Constâncio (Eds.)

TITLE
Essays on Values – Volume 3

EDITORS
Maria João Mayer Branco and João Constâncio

PUBLISHER
Instituto de Filosofia da Nova (IFILNOVA)
Faculdade de Ciências Sociais e Humanas
Universidade NOVA de Lisboa
ifilnova@fcsh.unl.pt
www.ifilnova.pt

GRAPHIC DESIGN
Filipe Pinto

PRINT
ARTIPOL – Artes Gráficas

ISBN
978-989-35468-2-6 (Electronic version)
978-989-35468-5-7 (Paperback)

DEPÓSITO LEGAL
529486/24

DOI
https://doi.org/10.34619/yocl-uo0e

FUNDING
This work is funded by national funds
through FCT – Fundação para a Ciência
e a Tecnologia under the project UIDP/00183/2020.

ESSAYS ON VALUES
Volume 3

IfILNOVA

ESSAYS ON VALUES

Volume 3

Editors

Maria João Mayer Branco and João Constâncio

Editorial Note

These three volumes, entitled *Essays On Values*, bring together forty-one recent articles by researchers at the Nova Institute of Philosophy (IFILNOVA). They are a small sample of everything that, in the last four years, the Institute's researchers have published, in English, in indexed journals and collections of essays with peer review. As a whole, they reflect very well the research work that is done at IFILNOVA.

Section I. of ***Volume 1*** gathers six articles that deal directly with the question "what are values?", the question that guides all the work of the institute's different laboratories and research groups. The first article, by Susana Cadilha and Vítor Guerreiro, results from work developed in the Laboratory of Ethics and Political Philosophy (EPLab); the second, by João Constâncio, from the Lisbon Nietzsche Group; the third, by Alexandra Dias Fortes, from the Lisbon Wittgenstein Group; the third and fifth, by Nuno Fonseca, and Maria Filomena Molder, from the Aesthetics and Philosophy of Art Group of the Laboratory of Culture and Value (CultureLab); the last, by Erich H. Rast, from the Philosophy of Language and Argumentation Theory Group and the Lisbon Mind, Cognition & Knowledge Group of the Laboratory of Argumentation, Cognition, and Language (ArgLab).

Section II. brings together three articles by members of the Lisbon Nietzsche Group. Since 2010, the Lisbon Nietzsche Group has completed several funded projects, and has established itself as a leading international research group on Nietzsche's thought. The three articles demonstrate the crucial importance of the question of values in Nietzsche's work, always thought from the perspective of the possibility of a "transvaluation of all values". Maria João Mayer Branco's article focuses on the value of introspection, and how Nietzsche anticipates Wittgenstein's "expressivist" view of the "the Peculiar Grammar of the Word 'I'" and the impossibility of private languages. Marta Faustino's article considers the theme of affirmation and the value of life through the interpretation of Nietzsche's reflection on truthfulness, intellectual honesty and courage in the light of Michel Foucault's work on *parrhesia*. Pietro Gori's article studies how Nietzsche creates a new anthropological ideal based on his enquiry into the values of the "good European".

The area of Wittgenstein studies has had a strong influence on the institute since the time when it was a philosophy of language institute. The Wittgensteinian distinction between facts and values was decisive in defining the question of values as the central issue of IFILNOVA's research project, replacing the focus on philosophy of language. More recently, the focus of research at the Lisbon Wittgenstein Group has been on epistemic values, in particular in their connection with the question of religious belief. In **Section III.**, Nuno Venturinha's article examines, in the light of an epistemological standpoint, the way Wittgenstein thinks about the possibility of translation. Robert Vinten's article argues that Wittgenstein's thought contains elements for a critique of the concept of justice and of the liberal political visions of both Richard Rorty and Chantal Mouffe, despite the fact that both have drawn inspiration from Wittgenstein. Benedetta Zavatta's article questions the value of mythology by thinking of it as a disease of language — not only in Wittgenstein, but also in a whole philosophical tradition that preceded him.

The existence of a research group in ancient philosophy is a recent but very promising development in the life of IFILNOVA. **Section IV.** includes two articles by members of the group. Paulo Alexandre Lima's article considers the critique of misology and the value of discourse in Plato's *Phaedo*. Hélder Telo's article examines the pedagogical and protreptic value of imperfection in Plato's work.

Section I. of ***Volume 2*** includes seven articles by researchers working on questions of aesthetics at CultureLab. Three of these articles, by Ana Falcato, Bartholomew Ryan and Tatiana Salem Levy, show how important the study of the relationship between philosophy and literature is at the Institute. Several of the CultureLab researchers investigate the possibility that the philosophical concept of "value" implies a transformation of lived values into objects of knowledge and instrumental calculation, and that literature, especially in authors such as Joyce or Coetzee, has always known how to avoid this kind of objectification. Bartholomew Ryan's article is also linked to that of Nélio Conceição. Both resulted from the research work carried out in the funded project OBRA — Fragmentation and Reconfiguration: the experience of the city between art and philosophy, coordinated by Maria

Filomena Molder and Nélio Conceição. The articles by Maile Colbert and Ana Godinho deal with questions concerning aesthetic values from the point of view of sound and drawing, respectively. João Lemos' article is a perfect example of the work that is done on Kant at the Institute, in particular on the relationship between aesthetic values and moral values.

Because film studies is a research area that mobilises a very significant number of researchers at IFILNOVA, it has been separated from the other research areas in Aesthetics for over ten years now, and is explored in an autonomous laboratory, CineLab. The articles in **Section II.** showcase the work that has been done in this area. The articles by Stefanie Baumann, Patrícia Castello-Branco, Paulo Stellino, Susana Nascimento Duarte and Susana Viegas reveal the importance of film studies for the research on fundamental authors in the history of philosophy, such as Kant, Adorno, Wittgenstein, Deleuze, or Foucault, but also the autonomously philosophical character of the works of fundamental authors in the history of cinema, such as Herzog, Straub/Huillet, Faroki or Manoel de Oliveira. The article by Gabriele De Angelis is the result of work carried out in the Ethics and Politics Laboratory (EPLab) but has been included in this section because it is an example of the intersection between laboratories of the institute, as it uses three films to discuss a crucial political issue of our time, the migration and refugee crisis in Europe.

IFILNOVA began as an institute for the philosophy of language. The question of values became the institute's central theme at a time when the philosophy of language was still the dominant area of study of the majority of its researchers. It was also at that time – around 2011 _ that the institute created the ArgLab and started to specialise in argumentation theory and mind and reasoning. ArgLab very quickly gained international recognition in this area. The articles in **Section I.** of ***Volume 3*** belong to this context. They all deal with Argumentation and Language. The article by Marcin Lewinski and Pedro Abreu and the article by Dima Mohammed and Maria Grazia Rossi mirror well the work developed by the institute in the area of argumentation and applied logic, in particular regarding the value issues raised by the COVID-19 crisis. The separate article by Maria Grazia Rossi is a case of the practical application of the theory of metaphor to the field of healthcare communication, a theme that has been heavily funded in projects carried out at the Arglab. The article by Giulia Terzian and Maria Inês Corbalán is emblematic of the intersection between linguistics and philosophy in the conceptual research about language.

The four articles in **Section II.** deal with questions concerning ethical and political values. Although from very different perspectives, the articles by Erik Bordeleau and Giovanbattista Tusa have in common a critique of capitalism and a questioning of its values. The discussion of political correctness in Filipe Nobre Faria's article and that of the concept of a People in Regina Queiroz's are investigations into the values of liberal democracies and how best to defend them.

The emotions, embodiment and agency are three themes of great importance in the work of several researchers at the institute. The link between these themes and the question of values is evident when one considers values as something that, far from being a mere abstraction or mental construct, is constitutive of the individual and collective life of human beings. The three themes are present in all the articles in **Section III.** The articles by Dina Mendonça and Robert W. Clowes have in common that they deal with the question of the depth of the mind. But the former approaches it from the perspective of the philosophy of emotions, the latter from the perspective of the philosophy of cognition. The article by Fabrizio Macagno, Chrysi Rapanta, Elisabeth Mayweg-Paus and Mercè Garcia-Milà deals with the concept of empathy as both an emotion and a value. The articles by António de Castro Caeiro and Luís Aguiar de Sousa reflect on the nature of the emotions, embodiment and agency in the light of the study of key moments in the history of Western philosophy: in the first case, the phenomenology of boredom in the work of Martin Heidegger; in the second, the metaphysics of Arthur Schopenhauer. Alberto Oya's article reflects on the nature and value of the religious experience. This article is published here for the first time, and so is Benedetta Zavatta's in volume one. Most articles in this collection have been originally published in Open Access journals, but some are republished here with the permission of the editors, to whom we are thankful.

Maria João Mayer Branco
João Constâncio

I.
Argumentation and Language

Arguing about “COVID”: Metalinguistic Arguments on What Counts as a “COVID-19 Death”

Marcin Lewiński and Pedro Abreu

Lewiński, M. & Abreu, P. (2022). Arguing About "COVID":
Metalinguistic arguments on what Counts as a "COVID-19 death".
In S. Oswald, M. Lewiński, S. Greco, & S. Villata (Eds.),
The pandemic of argumentation **(Vol. 43, pp. 17-41).**
(Argumentation Library; Vol. 43). Springer.
https://doi.org/10.1007/978-3-030-91017-4_2

1. Introduction

On March 11, 2020, the World Health Organization declared the COVID-19 epidemic rapidly spreading from China to most other countries in the world a "pandemic." A month later, on April 16, that same organization published *International Guidelines for Certification and Classification (Coding) of Covid-19 as Cause of Death, Based on ICD: International Statistical Classification of Diseases*. At that time, four months into the deadly first wave of infections, comparability of health and mortality data across all the affected countries became a key concern, as different countries seemed to be reporting and discussing different things. As a body mandated to protect international public health via, among other measures, a uniform classification of diseases, the WHO formulated the following "definition for deaths due to COVID-19":

> A death due to COVID-19 is defined for surveillance purposes as a death resulting from a clinically compatible illness, in a probable or confirmed COVID-19 case, unless there is a clear alternative cause of death that cannot be related to COVID disease (e.g. trauma). There should be no period of complete recovery from COVID-19 between illness and death.
>
> A death due to COVID-19 may not be attributed to another disease (e.g. cancer) and should be counted independently of preexisting conditions that are suspected of triggering a severe course of COVID-19. ("International Guidelines", p. 3)

We will return to these guidelines for further analysis in Section 3, but what is immediately striking about them is that they mix substantive and linguistic concerns to a puzzling effect. On the one hand, WHO is discussing and organizing substantive medical knowledge over "cause[s] of death" in "probable or confirmed COVID-19 case[s]" where, one would assume, the weight of scientific evidence is decisive. On the other hand, the organization presents its main results as a "definition" and "classification", which are two paradigmatic devices for metalinguistic and conceptual work. And this conceptual work is of paramount importance: "probable" COVID-19 cases are treated on a par with "confirmed" cases, and "independent" attribution of COVID-19 deaths is mandated even if other "preexisting conditions" such as cancer might have contributed to COVID-19 being severe enough to actually cause death. As is clear across the WHO's document and in the broader debate over the issue, these are neither scientifically determined nor arbitrary conceptual choices. Instead, in the cases we discuss below, reasonable even if characteristically inconclusive arguments are given to justify any such choice.

In this contribution, we explore the plausibility and consequences of treating such arguments as metalinguistic arguments. While unquestionably related to the epidemiological and public health issues, these arguments are also arguments about how a term should be used. As such, they touch upon some of the foundational issues in meta-semantics, discussed in the recent literature on metalinguistic negotiations, conceptual ethics, and conceptual engineering (Burgess, Cappelen, & Plunkett, 2020; Burgess & Plunkett, 2013; Cappelen, 2018; Plunkett, 2015; Plunkett & Sundell, 2013; 2021). Against this background, we analyze in particular how in the debate over what a COVID-19 death is, epistemic and practical reasons are intertwined in nuanced and complex ways to produce an interesting type of *metalinguistic interventions*.

We proceed as follows. In section 2 we provide the theoretical basis for our analysis. We introduce the phenomenon of what we summarily call *metalinguistic interventions*, present their three key features particularly relevant to our case, and offer distinctions instrumental in grasping the rather non-standard type of metalinguistic interventions related to "COVID-19 death." In section 3, we analyze official statements (of WHO, national governments) and media reports to critically reconstruct the metalinguistic elements of the dispute in terms of prevailing forms of argumentation used. In section 4, we discuss this analysis by developing two theoretically relevant points. First, the metalinguistic arguments revealed are inextricably linked to substantive, scientific issues and are partly determined by the imperfect character of our epistemic position on the subject. Second, they work in the service of broader practical arguments whereby scientific results are weighted against broader public policy values. We close by arguing that, in these ways, public metalinguistic arguments, while being a class of their own in need of precise analysis (see also Schiappa, 2003; Ludlow, 2014; Pruś, 2021), are of key importance to broader public debates and should be recognized as such.

2. Metalinguistic interventions

Metalinguistic uses of language have long been recognized as part and parcel of our communication. Perhaps most famously, Horn (1985) identified the mechanisms of "metalinguistic negation", a form of negation that is not a logical operator on truth-conditional propositions, but rather an objection to previous uses of language perceived as erroneous or infelicitous on grounds ranging from prosodic to conceptual. A good example of *conceptual* metalinguistic negation marked one of the twists in the public discourse over the COVID-19 pandemic. On September 26, 2020, Richard Horton, the editor-in-chief of *The Lancet*, one of the medical journals publishing peer-reviewed research instrumental to the scientific understanding of COVID-19, published a commentary (Horton, 2020) entitled:

(1) COVID-19 is not a pandemic.

This title, taken out of context, has instantly become a viral sensation for the negationist argument[1], thus turning Horton's well-intentioned conceptual "precisation"[2] into a perilous slogan for a standpoint he vehemently opposes (see Paglieri, 2021). But it takes only about 2 minutes to realize Horton's argument was impeccably metalinguistic:

(1a) COVID-19 is not a pandemic. It is a syndemic. [...] The notion of a syndemic [...] reveals biological and social

[1] As evidenced in the discussion on Horton's Twitter account immediately after the publication of the piece: https://twitter.com/richardhorton1/status/1309384015464587264?lang=en.

[2] For a discussion of various forms of "precising definitions" vis-à-vis Carnap's scientific "explication", see Brun (2016).

> interactions that are important for prognosis, treatment, and health policy. Limiting the harm caused by SARS-CoV-2 will demand far greater attention to NCDs [non-communicable diseases] and socioeconomic inequality than has hitherto been admitted. (Horton, 2020)[3]

As is clear in (1a), Horton's argument for conceptual shift from PANDEMIC to SYNDEMIC is justified on two grounds: scientific precision and public health response, with the latter taking the upper hand.[4] We will return to this interrelation of epistemic and practical arguments in our discussion below.

Such reasoned metalinguistic negations are, in our view, but one species of the argumentative and linguistic mechanisms that underlie public discussions where *metalinguistic intervention* (MI) plays a key role.[5] Attention to MI, encompassing various forms of reflection, discussion, and action on meanings, has been growing notably in recent analytic philosophy under various labels: *ameliorative analysis* (Haslanger, 2012), *conceptual engineering* (Cappelen, 2018), *conceptual*

[3] More precisely, this is an instance of a metalinguistic negation via the hypernym-hyponym relation ("Around here we don't LIKE coffee - we LOVE it"; "The wine is not GOOD, it's EXCELLENT"), discussed by Horn and others. The hypernym-hyponym relation can be given a scalar implicature interpretation: "One frequent use of metalinguistic negation – indeed, virtually universal (but cf. §5 below) – is as a way of disconnecting the implicated upper bound of weak scalar predicates." (Horn, 1985, pp. 139ff.).

[4] "[N]o matter how effective a treatment or protective a vaccine, the pursuit of a purely biomedical solution to COVID-19 will fail. […] Approaching COVID-19 as a syndemic will invite a larger vision, one encompassing education, employment, housing, food, and environment. Viewing COVID-19 only as a pandemic excludes such a broader but necessary prospectus" (Horton, 2020).

[5] It is important to stress here that throughout the chapter we use the term "metalinguistic" in a broad sense, as any explicit or implicit form of attempted intervention on the meanings of the expressions used. Some participants in the discussion on the issue—most notably Plunkett & Sundell (2013, 2021) and Ludlow (2014)—use instead "metalinguistic" in the specific sense of expressions that are implicitly used (rather than explicitly mentioned) not to communicate a fact but, assuming common knowledge of the facts, to communicate how these expressions should be used. As a result, for us explicit definitional disputes over, e.g., what counts as a COVID-19 death are thus metalinguistic, while in the narrower sense of Plunkett & Sundell they would rather be "canonical" disputes over which concepts to employ.

ethics (Burgess & Plunkett, 2013), *meaning litigation* (Ludlow, 2014), *metalinguistic negotiations* (Plunkett & Sundell, 2013, 2021; Plunkett, 2015), *verbal disputes* (Chalmers, 2011). While rooted in classic debates over the possibility of revisionary and pluralist approaches to meaning (Carnap, Quine, Davidson, Kripke, Putnam, Burge), this reinvigorated attention brings a new sense of relevance and urgency, as well as new methods, to the philosophical study of public uses of language. Lively theoretical disputes over the semantic/pragmatic nature of MIs, their metasemantic underpinnings, speakers' control over meaning, social and political functions of MIs, their potential for amelioration or perversion of meaning, permeate this literature (Burgess, Cappelen, & Plunkett, 2020; Marques & Wikforss, 2020). Still, the idea that MIs are often worthwhile and even central to public discussions is widely shared (see, however, Marques, 2017 and Stojanovic, 2012 for limitations).

An obvious objection to our approach would be to see the discussion over "pandemic" and "COVID-19 deaths" as basically a scientific dispute over facts. At the stage where the dispute takes place, we only have adequate epistemic access to a small fraction of the facts; we disagree about the rest because we infer different things about that rest based on the little knowledge we do share. For instance, in the case of COVID-19 deaths, the dispute revolves around different methodologies for calculating numbers of fatalities under fragmentary information, whereby full-proof medical evidence as to the causes of death of the thousands of suspected cases is missing. As a result, there is nothing metalinguistic patently involved just yet: after all, one of the defining characteristics of MIs is that disputants possess and mutually agree on all the relevant facts, and yet they disagree in virtue of the incompatible conceptual views they advocate on normative grounds (Ludlow, 2014; Plunkett, 2015; Plunkett & Sundell, 2013, 2021; Schiappa, 2003). They thus fix their beliefs, while trying to solve for the meaning. Accordingly, this objection would maintain that until any forthcoming empirical facts might be decisive in adjudicating the dispute, it is essentially a substantive, ground-level dispute.

This objection can be resisted on two grounds. First, it assumes that there is, eventually, the scientific truth of the matter on what a COVID-19 death is, and that the problem lies in the scarce resources and underdeveloped methods to arrive at that truth (e.g., precise and massive tests and autopsies). But this assumption can be legitimately challenged: multiple notions and conceptions of *cause* and, more specifically, *cause of death* have played a role in various scientific and medical contexts (Clarke & Russo, 2016; Lindahl, 1988, 2021; Reiss, 2016; Reiss & Ankeny, 2016). It is all but clear that any single one of these should or could be elected as *the* right or privileged one with which to form a univocal scientific concept of COVID-19 DEATH. Additionally, there is the issue of numerous particular cases of especially indeterminate nature, even within what seems to be a fixed framework.[6] Lindahl (2021) gives the example of situations of COVID-19 infection in patients with cancer, in which the two diseases "reciprocally interact, increasing the seriousness of the outcome" (2021, p. 72), thus rendering dubious the possibility of a clear choice of either morbidity as *the* underlying cause of death. Indeed, one can claim that "the cancer and the COVID-19 *jointly* initiated the train of morbid events leading directly to death" (Lindahl, 2021, p. 72, italics in the original), and given that only one can be reported on the death certificates, discretionary decisions need to be made by coroners. That's where the guidelines such as the ones of WHO come to the rescue: complex situations of a rapidly spreading pandemic driven by a hitherto unknown virus are rife with uncertainty, indeterminacy, and certain arbitrariness of results that cannot be conclusively overcome by scientific means alone for the purposes of concerted public health response.

That brings us to the second argument against the objection. Even if the assumption of the scientific truth of the matter proved to be at least approximately adequate (perhaps with better diagnostic methods being developed and widely implemented), for our argument to take off the

[6] There is a debate among medical practitioners over the accuracy of that cause of death reporting in COVID-19 patients. The problem is well exemplified by the Swedish study of Nilsson et al., 2021: "Death in home healthcare during the first pandemic wave mostly affected individuals already vulnerable due to severe frailty and very advanced age. In this group of subjects, COVID-19 was assessed as contributing to death in two-thirds of the individuals, and less frequently, it was the dominant cause of death (13%). One of every five individuals was assessed as dying from another cause than COVID-19" (Nilsson et al., 2021, p. 3). But even the studies that claim reporting is indeed accurate within the national and international (WHO) reporting guidelines (e.g., Elezkurtaj et al., 2021; Slater et al., 2020), are not immune to the deeper problem of the indeterminacy of the cause of death we discuss here.

ground we do not need to resist this objection so far as the SCIENTIFIC CONCEPT[7] is concerned. Indeed, the objection can help us make clear that there are (at least) two concepts and two sets of issues converging and being conflated in these discussions. It does so by rendering it clear that the scientific concept and a set of related issues constitute just part of the concerns of health authorities when they discuss and issue operational definitions. On the other hand, at the same time, there is also a different concept— the INSTITUTIONAL CONCEPT — and a set of issues that are fully determined by institutional declarations. International and national health authorities, facing the need for urgent and decisive action under uncertainty, propose, discuss and establish uniform and operationally precise "definitions" and "classifications" which make it possible to overcome remaining uncertainties. It is the metalinguistic interventions on this second concept that we focus on here.

Our focus on institutional concepts as the domain of MIs over COVID-19 deaths is inspired by Searle's social ontology (Searle, 1995, 2010).[8] Within this theoretical framework, by declaring a given epidemiological situation a "pandemic", the WHO creates a new institutional reality in which various institutions and agents are endowed with new rights and obligations. For instance, we have the right to resort to the *force majeure* clause to cancel or alter our obligations and, simultaneously, we have the obligation to follow strict health-related regulations and limitations (e.g., travel bans). These conditions make up the declarative status of these acts. *Declarations* are precisely the speech acts that create new social realities by the very fact of being felicitously

[7] For the purposes of this chapter, we stick to the prevalent (even if mildly sloppy) practice of using 'concepts' and 'meanings' interchangeably so as to signal our neutrality on the questions concerning the nature of our representational devices. While this is largely inconsequential to our arguments here, we are well aware of the ongoing dispute over this practice (see Eklund, 2021; Machery, 2009; Sawyer, 2018, 2020).

[8] While Cappelen (2018, pp. 44-46) briefly discusses Searle's social ontology as an approach which potentially affords revision and amelioration of concepts that are constitutive of social facts, he doesn't explore this connection any further. Like us, Schiappa (2003) also draws attention to Searle's realm of institutional facts and advocates that one appropriate form of definition is "X counts as Y in context C", but, similarly to Cappelen, treats this connection rather perfunctorily. Otherwise, Schiappa offers a framework much more resolutely constructionist than we find necessary and justifiable.

performed: a declaration of war by a legitimate head of state just starts the war, an official announcement of firing an employee by the employer just is firing him, etc. (Searle, 1975, 1995, 2010). All the same, there is a special sub-type of declarations that still create institutional facts but are grounded in some natural or social facts, namely, *representative declarations* (Searle, 1975, pp. 360-361): a judge declaring someone guilty just makes this person guilty, and yet also makes a factual statement to the effect that the accused actually did commit such-and-such criminal acts. Similarly, the WHO declares a "pandemic" because, to the best of WHO's knowledge, there actually is a pandemic.[9] There are, then, belief-relevant sincerity conditions related to such acts that do not exist in pure declarations, e.g., in the act of declaring a war or opening an academic conference. As a result, one can be right or wrong in such declarations, and one can lie in them too.[10] Interestingly, in the case of correct declarations, the objective natural or social facts are coextensive with the declared institutional facts. However, in the case of incorrect or even manipulative declarations we have two parallel facts running their own course. For instance, for all the legal intents and purposes, we might act, and even be obliged to act, under the conditions of pandemic as an *institutional* fact, while the pandemic as an *epidemiological* fact is actually not happening (and vice versa, as witnessed by the situations where authorities declare an end to lockdown restrictions without obvious changes in epidemiological facts). There exist erroneous verdicts.

[9] See: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020: "WHO has been assessing this outbreak around the clock and we are deeply concerned both by the alarming levels of spread and severity, and by the alarming levels of inaction. We have therefore made the assessment that COVID-19 can be characterized as a pandemic. Pandemic is not a word to use lightly or carelessly. It is a word that, if misused, can cause unreasonable fear, or unjustified acceptance that the fight is over, leading to unnecessary suffering and death."

[10] In Searle's well-known terminology, for all declarations "the direction of fit is both words-to-world and world-to-words because [...] the performance of a declaration brings about a fit by its very successful performance" (1975, pp. 359-360). However, representative declarations have an additional words-to-world dimension characteristic of assertions. In this way, Searle is refining Austin's (1962) original class of truth-relevant "verdictives" as distinguished from pure "exercitives."

Note that this is precisely Horton's argument: the WHO declared the wrong kind of health emergency. Instead of 'pandemic', we should officially talk about 'syndemic', a concept that not only better captures the evolving epidemiological facts, but also points to more adequate ways of addressing the short- and long-term effects of COVID-19. SYNDEMIC is thus epistemically more precise and prescriptively more fruitful, thus meeting two classic criteria for conceptual work (Carnap, 1950; Brun, 2016; Dutilh Novaes, 2020; Plunkett, 2015).

Further, and most importantly to our discussion: in the case of pandemic, the WHO used their recognized prerogative to apply the standing declaration to an individual case at hand. *Standing declarations* are constitutive rules determining what would be an acceptable applied declaration (Searle, 2010, p. 13). In our case, it is within WHO's powers to declare a pandemic antecedently defined as "the worldwide spread of a new disease" – and they did just that on March 11, 2020.[11] However, one can also discuss and institute a standing declaration in the first place, thus fixing the general rule *X counts as Y in C*. This type of declaration takes the form of an institutional definition, or a part of it: e.g., *Dying with recognizable COVID-19 symptoms (dry cough, fever) but without any further evidence counts as dying of COVID-19 in the context of Belgian elderly care homes residents*. Institutional definitions, while linguistic, thus require an extra-linguistic institution, against Searle's arguments to the contrary (1975, p. 360; 2010, Chs. 4-5). Any such definition, when duly approved and recognized, becomes a standing declaration which, whenever implemented, creates an institutional fact, a recognized status that comes with certain rights and obligations, as described above.

While Searle's original intention was to theorize how institutional reality is constructed and maintained, we re-use his distinctions in order to precisely delineate the domain where metalinguistic arguments over "pandemic" and "Covid-19 death" – and multiple other similar cases – can and do happen. Given the relevance of epidemiological and medical facts, arguers are bound to discuss and pronounce *representative*

[11] https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

declarations. Further, since the focus of these arguments is not merely on how to apply a given concept under specific circumstances but rather how to "define" or "classify" that concept in the first place, the MIs concern *standing* declarations. Such standing representative declarations are the object of our study here.

In the literature, the complex interrelations between metalinguistic and substantive issues are well recognized (Chalmers, 2011; Plunkett, 2015). In principle, there might exist criteria for distinguishing between the two; in typical cases, we enter the realm of MIs when speakers continue to disagree while, factually speaking, all is said, done, and mutually agreed on (including facts over the other speakers' meanings).[12] However, most curious in the philosophical debates are various hybrid and messy cases. Indeed, the standing representative declarations we analyze in this chapter are clear and interesting instantiations of such mixed phenomena. Here, the declarative, definitional element accounts for the metalinguistic or conceptual aspect, while the representative element accounts for the substantive aspect. The dispute is thus indexed and accountable to some external reality—just as much as a judge's decision to declare someone guilty is—but, once a declaration is issued, it does become an institutional fact itself.

Beyond this fundamental aspect of conceptual work over institutional facts, we point to three key elements of the philosophical dispute over MIs, particularly relevant to an analysis such as ours.

First, as elaborated in their unique ways by Haslanger (2012), Plunkett & Sundell (2013, 2021; Plunkett, 2015), and others (e.g., Ludlow, 2014; Schiappa, 2003) MIs – or at least those most persistently argued about – are driven by *normative*, rather than descriptive, concerns. Plunkett & Sundell (2013, 2021) distinguish between *descriptive* metalinguistic disputes over how a term *is* used (e.g., "For us in Europe 'football' means a different game than for you guys in the USA") and *normative* metalinguistic disputes over how a term *should be* used (e.g., "Waterboarding is torture"; "Horses are athletes"). In contrast to the descriptive cases, the issue cannot be conclusively

[12] See Soria Ruiz (2021) and Stojanovic (2012) for further discussion on the distinction between metalinguistic and evaluative disputes, which *prima facie* share some of these features.

settled by appeal to current usage or some linguistic authority (e.g., by restriction to the current regulations within the legal domain) – which makes them disputes of particular philosophical interest. Plunkett & Sundell call them, somewhat misleadingly, metalinguistic 'negotiations' (see below). Normativity itself is, however, another forbiddingly complex notion.[13] For the sake of illustrative simplicity, we can divide normative grounds of MIs into three large classes, recognized since antiquity: the true, the good, and the beautiful. The former two are especially relevant to our discussion here. In the first place, one engages in MIs for the sake of epistemic enhancement. In doing so, one can appeal to Carnapian values of specifically scientific exactness, simplicity,[14] and fruitfulness in the pursuit of methodic inquiry (Carnap, 1950; Brun, 2016; Dutilh Novaes, 2020) or to a broader metaphysical value of "carving reality at its joints" (Sider, 2012; Scharp, 2020). Such appeals can support and explain prototypical examples of conceptual refinement such as those concerning FISH and ATOM (Carnap, 1950; Dutilh Novaes, 2020; Rast, 2020). In the second place, MIs work in the service of ethical concerns, that can be quite general and abstract or more applied, focused on concrete cases (Burgess, Cappelen, & Plunkett, 2020). Ideals of fairness, equality, or dignity are thus often invoked in attempts to intervene metalinguistically on a concept such as FREEDOM or on a concept such as MARRIAGE or TORTURE. Importantly, in either case, a broadly pragmatic approach can be defended, tying the grounds and forms of MIs to the goals at hand, e.g., those of scientific inquiry or of public policy. Plunkett & Sundell (2021) stress the primacy of such overarching practical goals when they insist that "arguing about whether waterboarding is torture is a way of arguing about *whether we should waterboard*, or about *how*

[13] Given that meaning itself can be considered a normative notion, thus encompassing descriptive disputes, one needs to further distinguish between normativity *internal* and *external* to the use of language. It is the latter type that is relevant here. Finally, merely *procedural*, minimal normativity in the sense of any rule-governed behavior vs. value-based *substantive* normativity should be distinguished. Again, it is the latter type that is relevant here. See Plunkett & Sundell (2013, 2021) and Plunkett (2015).

[14] Note that simplicity and similar notions such as elegance or parsimony are often considered "aesthetic values" in scientific theories, thus pertaining to the class of the beautiful. See, e.g., Ivanova (2017).

we should treat people that do it, or some other normative issue" (p. 162, emphasis in the original).[15]

As we shall see, these concerns are indeed crucial in the public debate over COVID at large, and COVID deaths in particular.

Second, MIs can be performed via disputes over terms and concepts explicitly *mentioned* as arguable, or via disputes over terms and concepts implicitly *used* as arguable (Burgess & Plunkett, 2013; Plunkett & Sundell, 2013, 2021; Rast, 2020). While the latter seem finer and more elusive thanks to their intricate pragmatic mechanisms, the former are more directly amenable to the analysis of the arguments driving the dispute. In this case, however, we would rather not call them metalinguistic 'negotiation' but 'argumentation', given the centrality of "rational conflict" to the concept (Plunkett, 2015): rational conflict, or a disagreement instigated by rational concerns, when managed on rational grounds via linguistic exchange, just *is* argumentation on the most standard meaning of the term (see van Eemeren & Grootendorst, 2004; Dutilh Novaes, 2021). Accordingly, one of the tenets of argumentation theory is that it is public argumentation, and not private reason, that promotes rationality, precisely due to its explicitness. For Johnson (2000), argumentation is not only rational, but *manifestly* rational, so that arguers can mutually see, test, and acknowledge the rationale behind inferential steps taken.[16] By contrast, 'negotiation' denotes a linguistic activity of arriving at a reciprocally agreeable private

[15] Responding to Cappelen's challenge that the dispute over whether waterboarding is torture is an object-level and not a metalinguistic dispute, that is, it is "about torture, not 'torture'" (Cappelen, 2018, p. 175), Plunkett & Sundell claim that "in many cases, the debate that really matters is not about the word 'torture' or about torture. It's about *waterboarding*, and whether we should be doing it." (2021, p. 162, emphasis in the original).

[16] "It is not just that the participants [in argumentation] embrace rationality, which they might do secretly but not publicly. No, the participants in the practice exhibit what it is to be rational. To give reasons; to weigh objections; to revise over them or to reject them – all of this describes a vintage performance of rationality. The arguer acknowledges that there are objections and problems with the position [...]. The critic acknowledges that there is rationality in the arguer's position." (Johnson, 2000, pp. 162-163). Pragma-dialectical "meta-theoretical principles" of *externalization* of commitments, and of *socialization, functionalization and dialectification* of argumentation similarly reinforce the link between explicitness and rationality of argumentation (van Eemeren & Grootendorst, 1984, 2004).

compromise, that *can* be rational, but can also be purely transactional (Godden & Casey, 2020).

In this way we second Ludlow's idea that the driving force behind MIs is to come up "with progressively more serviceable modulations via a normatively constrained process of argumentation" (Ludlow, 2014, p. 111). These processes can be based on analogical arguments or arguments from authority (Ludlow, 2014) or on various other forms of definitional and semantic arguments (for a recent overview, see Pruś, 2021). However, as we show below, in line with point one discussed above, practical reasoning seems to be a central type of argumentation grounding MIs.

Finally, one of the key concerns in conceptual ethics or engineering is this: can we really control the change of our concepts? Ludlow (2014) and linguists working within lexical pragmatics (Allott & Textor, 2012; Hall, 2017; Wilson, 2003) argue that in communicative contexts, speakers can tweak meanings via pragmatic or semantic modulations. For Ludlow, this idea comes with radical contextualism whereby interlocutors, as it were, create their "micro-languages" from scratch in any given conversational context, and thus are free to adjust their meaning at will. Diametrically opposed to this position, and rooted in a particularly unrelenting understanding of semantic externalism, we find Cappelen's lack of control argument: given that meanings (intensions and extensions) supervene on long-term patterns of usage within a broad linguistic community, local and individual attempts at meaning change can only have a minute and unpredictable impact, if any at all. Cappelen admits, however, that attempts at MIs continue, driven by normative concerns: even if a lasting, widespread semantic change is arguably beyond speakers' control, we still engage in MIs if only because "in general, we don't make normative judgments [...] only when we have worked out a strategy for how to change the world" (Cappelen, 2018, p. 75). Our normative reasons, discussed above in points 1 and 2, thus prevail over practical limitations: we pursue, however unwittingly, the "right" meanings of our words even if we cannot fully understand, let alone control, processes of meaning change.

In between these two extremes, various options for effective intervention on our concepts and meanings are conceivable and have been explored in the literature, from forms of metalinguistic activism (Sterken, 2020), to engagement in "collective long-range" meaning change efforts (Koch, 2021), or even the engineering away, from our very concepts of CONCEPT and MEANING, of whatever features stand in the way of agents' control over their representative devices (Riggs, 2019). While we are not in a position to further explore here, let alone resolve, this debate, we note a special context where control over meanings is well possible, and even expected. This is exactly the area of social ontology, discussed above. It is within the deontic powers of certain certified bodies – international organizations, constitutional assemblies, parliaments, municipal and faculty councils, but also courts of various instances, notably supreme courts – to declare on certain conceptual choices via their legal authority to do so, thus pronouncing binding semantic resolutions. Vivid examples of this – anything from what is a PERSON to SUSTAINABLE FASHION to SANDWICH – are discussed by Ludlow (2014) and within argumentation theory (Schiappa, 1993; 2003; Greco & De Cock, 2021). In such instances, Searle's formula for constitutive rules operative in declarative speech acts – *X counts as Y in context C* – replicates itself thus creating social reality, with its network of intentional states and background capacities (Searle, 1995, 2010).

In this way, we thus carved out our approach to MIs: we specifically focus on MIs 1) grounded in various forms of *normative argumentation*, 2) *explicitly* debatable *in the public sphere* and 3) aimed at *meaning change in the domain of institutional facts*. These three characteristics jointly converge on an approach to MIs particularly fruitful in our inquiry over what counts as a COVID-19 death.

3. Arguing over what a COVID-19 death is

In this section, we argue that public understanding of the COVID-19 pandemic, and a successful response to it, depend in part on an answer to a seemingly simple question: What do or should we mean by a "COVID-19 death"? This concern is reflected in the metalinguistic arguments of health authorities and public media that we analyze here.

3.1. The early confusion

Consider the discussion over case mortality rates of COVID-19 that, in the early stages of the pandemic in Europe (February-March 2020), varied from 1% (Germany) to 10% (Italy, Spain, Belgium). Explanations abounded on how to account for this difference.[17] Obviously, "facts on the ground" were brought up: demographics such as average population age, health, and density; overall quality of healthcare with a focus on available ICU beds and ventilators; government response, including the timing and severity of the lockdown measures; availability of the personal protective equipment (masks, gloves); even air quality. Further, testing methodology was recognized as playing a key role: tests could be limited to patients with severe symptoms and their direct contacts, resulting in higher mortality rates reported, or included a broader, asymptomatic population, producing lower rates. Quite recognizably, such background facts and methods are two standard grounds for substantive disputes over this and similar cases.

However, from the onset of the pandemic, a third line of explanation has been present, one that focuses on the "differences caused by clinical definitions of what counts as a Covid-19 death" ("BBC report").[18] Such differences can be seen as particularly artificial when urgent and concerted action demand adequate worldwide comparison and coordination in the counting of cases. As we have already mentioned above, the question of "what counts as a Covid-19 death" does not admit of an obvious, single answer. Given the virus has been particularly lethal among older patients with other underlying illnesses (so called "comorbidities"), how were doctors advised to discern whether a patient died "as a result" of COVID-19, or rather a bacterial pneumonia, terminal cancer, or heart attack? While during

[17] See, e.g., https://www.bbc.com/future/article/20200401-coronavirus-why-death-and-mortality-rates-differ
https://www.theguardian.com/world/2020/apr/24/is-comparing-covid-19-death-rates-across-europe-helpful-

[18] https://www.bbc.com/future/article/20200401-coronavirus-why-death-and-mortality-rates-differ.

the early stages of the pandemic most countries instituted a simple principle—any death of a patient tested positive "counts as" a death "caused by" COVID-19—actual clinical practice across and within different European countries varied, spurring a dispute among health professionals, policymakers, and the general population.[19]

Here, we defend the position that this problem—as well as any of the attempted or possible solutions—is a metalinguistic one.[20] Some institutions we analyze below explicitly mention this as being a matter of *definitions* and *classifications* (WHO, ONS in the UK). However, even more importantly, a confirmation that the relevant lack of coordination in accounting for COVID-19 deaths is, at least in part, semantic in nature stems from the fact that it can straightforwardly give rise to verbal and metalinguistic disputes. It is quite natural, in this context, for someone to abstain from answering an object-level question like "Is this a COVID-19 death?", or "Did x die of COVID-19?", and to reply, instead, at the meta-level, with something like "It depends on what you mean by 'COVID-19 death'."[21]

Indeed, in the spring of 2020, nascent metalinguistic arguments began to emerge. The predominant line defended the broad definition as an adequate indicator of the dangers of the pandemic. Others called

[19] For a representative example of arguments in this early dispute, see the Ioannidis-Taleb debate analyzed in Antiochou & Psillos (2022).

[20] Note that in claiming that the issue is of a metalinguistic nature, we don't take ourselves to be committed to its not being also substantive. Despite its pragmatic usefulness, we are generally suspicious of the possibility of a principled, clear, and robust distinction between verbal (meta-level) and substantive (object-level) issues, disputes, and arguments. This is not the place to elaborate on this topic. We present further details of this view in a forthcoming article.

[21] Soria Ruiz (2021) formulates three helpful tests for ascertaining the metalinguistic character of a given dispute. These tests further support our arguments, as the differences in counting something as a COVID-19 death indeed share the relevant properties with other paradigmatic metalinguistic disputes, namely: 1) consider-embeddings of the disputed expression are felicitous, e.g., "WHO considers this to be a case of COVID-19 death (while gov. uk doesn't)"; 2) non-ironical/humorous metalinguistic comparatives appear perfectly possible in the relevant contexts, e.g., "This is more a COVID-19 *related* death than simply a COVID-19 death"; 3) finally, in numerous such cases, the most salient question under discussion is precisely the metalinguistic one: "What should count as a COVID-19 death?"

for a more precise, narrower approach needed for better clinical practice and public response: COVID deaths need to be *actual* COVID deaths, not just deaths of people who happened to have a positive result, but in fact died from other illnesses, or simply old age.[22] In an apt rejoinder, the liberal side responded that, given the early scarcity of tests, counting only the positively tested cases amounted to a gross underestimation of the actual scope of the pandemic.[23] Compared to other pressing epidemiological concerns, this might sound as futile verbal disputes. Still, these semantic arguments illustrate the first efforts to understand and fix what 'COVID-19 death' means and to properly gauge the impact of the pandemic across the world's population. An argument from analogy was also put forth (see "BBC Report"): In the case of the 2009 swine flu pandemic, depending on the way health professionals "assigned causation", the death rate varied from dangerous 5.1% (early reports) to mere 0.02% (current corrected rate, based on a careful revision of medical data, including definitions and assignments registered in death certificates). Should the disputes over COVID-19 reveal a similar effect, then arguments over meaning would be very much worth having. Yet, as the pandemic raged in the spring of 2020, no consensual and conclusive reasons managed to decisively tilt these meaning disputes toward one solution or another. At this stage, international and national institutions stepped in.

3.2. Solution 1: WHO's broad concept

In April 2020, the WHO intervened, producing "International guidelines for certification and classification of COVID-19 as cause of death based on ICD: International Statistical Classification of Diseases."[24] Referring

[22] In the words of a Belgian virologist Marc Van Ranst: "It now seems that people are only dying of COVID-19 in our nursing homes, while there are deaths there even in normal times, given the very high average age of their residents." As quoted in https://www.politico.eu/article/why-is-belgiums-death-toll-sohigh/

[23] For various reasons, see here: https://www.healthdata.org/research-analysis/diseases-injuries/covid/estimation-total-and-excess-mortality-due-covid-19

[24] https://www.who.int/docs/default-source/classification/icd/covid-19/guidelines-cause-of-death-covid-19-20200420-en.pdf

to "probable or confirmed COVID-19 cases" WHO's "definition for deaths due to COVID-19" stipulated that:

> (2) A death due to COVID-19 is defined for surveillance purposes as a death resulting from a clinically compatible illness, in a probable or confirmed COVID-19 case, unless there is a clear alternative cause of death that cannot be related to COVID disease (e.g. trauma). There should be no period of complete recovery from COVID-19 between illness and death.
>
> A death due to COVID-19 may not be attributed to another disease (e.g. cancer) and should be counted independently of preexisting conditions that are suspected of triggering a severe course of COVID-19. (p. 3)

The definition is lax in its epistemic demands and broad in its reach. Quite surprisingly, it counts merely "probable" cases on a par with "confirmed" ones and determines that deaths due to COVID-19 "should be counted independently of preexisting conditions", even those "preexisting conditions that are suspected of triggering a severe course of COVID-19" (for further discussion, see Amoretti & Lalumera, 2021 and Lindahl, 2021).

In view of our foregoing discussion, important questions arise: Can the WHO determine what a 'COVID-19 death' means? And, in this particular case: Did the WHO at least produce a sound argument for what a 'COVID-19 death' means or should mean?

Resorting to the distinctions introduced earlier, the WHO argues over an *institutional* concept of COVID-19 DEATH, precisely because of the obstacles to the deployment of a *scientific* concept, both principled (the nature of cause of death) and practical (insufficient capacity to test and perform autopsies). Not only is the very possibility of electing a single, natural, scientific concept of COVID-19 DEATH doubtful for the reasons discussed above, such a concept would not, in any case, be immediately adequate in the context of turmoil, fragmentary information, and pressure for quick measures and pronouncements. Under these circumstances, the WHO is able, indeed obligated, to intervene and fix the institutional (and operational) meaning for 'COVID-19 death'. It undoubtedly has the effective power to implement worldwide changes in how the term is applied in official documents and statements. Accordingly, the document starts with the broad standing declaration of what should count as COVID-19 death (2), and then moves on to specific instructions on how to apply this declaration in concrete cases (2a). It thus first offers an argument *to* definition and, once this is settled, an argument *from* definition (see Pruś, 2021; Rigotti & Greco, 2019). Importantly, the conceptual argument to definition is grounded in normative concerns "of importance for public health" that are relevant "for surveillance purposes" and "the most useful cause of death statistics possible." The concern for producing data "comparable to data from other countries" further reinforces this argument.

In this way, as also noted by Amoretti & Lalumera (2021) and Lindahl (2021), values other than medical or scientific accuracy govern this intervention. The WHO is explicit about the heterogeneity of considerations shaping their definitions and instructions:

> (2a) With reference to section 4.2.3 of volume 2 of ICD-10, the purpose of mortality classification (coding) is to produce the most useful cause of death statistics possible. Thus, whether a sequence is listed as 'rejected' or 'accepted' may reflect interests of importance for public health rather than what is acceptable from a purely medical point of view. Therefore, always apply these instructions, whether they can be considered medically correct or not. Individual countries should not correct what is assumed to be an error, since changes at the national level will lead to data that are less comparable to data from other countries, and thus less useful for analysis. (pp. 8-9)

Key scientific values such as precision and self-correction are thus overridden by a straightforward practical argument: in the current situation marked by scientific uncertainty and lack of consistency, and given our institutional mandate of protecting public health in an internationally coordinated manner, the best definition-qua-rule we can institute is: *any death resulting from a clinically compatible illness, in a probable or confirmed COVID-19 case, counts as COVID-19 death (unless there is a clear alternative cause of death that cannot be related to COVID disease) in the context of the current pandemic.*

3.3. Solution 2: Belgium's broad concept

A much-debated version of the WHO's definition of what counts as a COVID-19 death was introduced in Belgium. The controversy revolved around how Belgium decided to fix the meaning of 'COVID-19 deaths' by including in it 'probable deaths' and counting such cases in the official statistics of COVID-19 deaths.[25]

> (3) "As in other European countries, there wasn't enough test capacity in the beginning to extensively test patients in nursing homes," said Joris Moonen, a spokesperson for the agency that oversees nursing homes in the Dutch-speaking region Flanders. "We choose to register every death who had potentially died from COVID-19 to detect in which nursing homes the virus had hit. We were aware this would lead to an overestimation but found the signaling more important.[26]

Placing the value of public health "signaling" over the possible epistemic "overestimation" mirrors WHO's arguments. Critics called it simply "stupid", on both epistemic and practical grounds. First, as reported, "of Belgium's registered deaths, 44 percent died in hospital (and were tested). The majority 54 percent died in a nursing home — and only in

[25] See https://www.politico.eu/article/why-is-belgiums-death-toll-so-high/; https://www.politico.eu/article/in-defense-of-belgium-coronavirus-covid19-pandemic-response/; https://www.nytimes.com/2020/08/08/world/europe/coronavirus-nursing-homes-elderly.html; https://www.theguardian.com/world/2020/apr/24/is-comparing-covid-19-death-rates-across-europe-helpful-

[26] https://www.politico.eu/article/why-is-belgiums-death-toll-so-high/

7.8 percent of those cases was COVID-19 confirmed as the cause."[27] That leaves almost 50% of official numbers in the medical dark. Second, such approach possibly had adverse practical consequences: "Neighboring countries may be less likely to reopen their borders for Belgian companies or tourists once European governments start to loosen their confinement measures."[28] Indeed, in the early weeks of the pandemic (March-April 2020), Belgium had the highest per capita death rates in Europe and even in the world. This indicates another set of pressing practical arguments relevant in debating the institutional concept of COVID-19 DEATH, namely, public image factors (e.g., not appearing a failed state) and commercial interests (of national businesses or tourists).[29]

However, defenders of the government's policy produced counter-counter-arguments:[30]

> *(3a) "It's important that people are aware of the deceases outside the hospitals," Van Gucht [who chairs the government's scientific committee for coronavirus] said. "A broad way of counting enables us to monitor and quickly intervene where needed. Numbers are very important to create a sense of urgency — for example for the nursing homes. Belgium shouldn't be ashamed about that."*[31]

Again, the argument of efficient public health response takes precedence here over slow-paced medical accuracy. Importantly, while Belgian officials explicitly discuss "a way of counting", it is worth noting that in this context the *statistical* sense of "counting" is derivative of the *definitional* sense of "counting" as in the formula X "counts as" Y in context C. This, as we have argued, accounts for the metalinguistic aspect of the dispute over institutional facts.

[27] https://www.politico.eu/article/why-is-belgiums-death-toll-so-high/.

[28] https://www.politico.eu/article/why-is-belgiums-death-toll-so-high/.

[29] We thank an anonymous reviewer for pressing this point.

[30] See also the official defense of Maggie De Block, Belgium's minister of public health: https://www.politico.eu/article/in-defense-of-belgium-coronavirus-covid19-pandemic-response/;

[31] https://www.politico.eu/article/why-is-belgiums-death-toll-so-high/.

3.4. Solution 3: UK's narrow concept: ONS vs. GOV.UK

Belgium's chief scientist's argument that "it's important that people are aware of the deceases outside the hospitals" is not a standalone reason, but rather a direct objection to the decisions taken in other European countries, notably the United Kingdom, the recently estranged ex-member of the European Union. In the UK, the government instituted a principle that only deaths 1) with a confirmed positive COVID-19 test and 2) those occurring in hospitals count as COVID-19 deaths to be reported in official statistics. This practice directly contradicted WHO's instructions and practices of countries such as Belgium. Unsurprisingly, this triggered a public debate, outside and inside of the UK. A useful summary of this early debate can be found in the official blog of ONS, the Office for National Statistics:[32]

> (4) ONS figures by actual date of death (death occurrence) tend to be higher than the GOV.UK figures for the same day. This is because:
>
> - We include all deaths where COVID-19 was mentioned on the death certificate, even if only suspected: the GOV.UK figures are only those deaths where the patient had a positive test result
> - We include deaths that happened anywhere in England and Wales, for example some might be in care homes: the GOV.UK figures are only those that happened in hospital.
>
> **So who is right about the number of deaths?**
>
> The issue is not really about right or wrong, but about each source of data having its own strengths and weaknesses.

[32] For further analysis of the COVID-19 debate in the UK, see Fairclough (2022).

> The figures published on GOV.UK are valuable because they are available very quickly, and give an indication of what is happening day by day. Their definition is also clear, so the limitations of the data can be understood. But they won't necessarily include all deaths involving COVID-19, such as those not in a hospital.
>
> Numbers produced by ONS are much slower to prepare, because they have to be certified by a doctor, registered and processed. But once ready, they are the most accurate and complete information.
>
> Using the complete death certificate allows us to analyse a lot of information, such as what other health conditions contributed to the death.[33]

This post nicely captures the institutional dilemmas to be resolved. UK Government has a "definition" of COVID-19 death that is clear, fast, and frugal. But ONS deems it too far removed from "the most accurate and complete information", something ONS is after in their approach. The government was responsive to such arguments, and in August 2020 changed its definition by removing the condition of *hospital* death, thus defining COVID-19 deaths as:

> (4a) deaths in people with COVID-19 that occur within 28 days of a first positive laboratory-confirmed test.[34]

Hence the condition of a "positive laboratory-confirmed test" remains necessary. In their justification of this decision, the government argued the following:

[33] https://blog.ons.gov.uk/2020/03/31/counting-deaths-involving-the-coronavirus-covid-19/

[34] https://publichealthmatters.blog.gov.uk/2020/08/12/behind-the-headlines-counting-covid-19-deaths/

(4b) ONS reports deaths where a doctor suspects COVID-19 as a cause – these data include a clinical assessment as recommended by WHO but are subject to variation in clinical judgement as to the cause of death.

In other words, the institutional extension of 'COVID-19 deaths' should be a subset of the scientific extension: discretionary powers of individual doctors, which inevitably include subjective suspicion and varied judgment, should not yield to "laboratory-confirmed" truth of the matter.[35] Both the WHO and the ONS are thus mistaken in their approach – and so is the Belgian government. Whichever way the argument goes, however, the British case demonstrates the possible transience of conceptual interventions, an issue central to Ludlow's (2014) framework. Certain conceptual solutions might be adequate in a certain context, while certain specific conditions hold, and inadequate when something changes. This consideration brings us to the last option for conceptualizing COVID-19 deaths in the context of, by mid-2021, a prolonged, unrelenting pandemic.

3.5. Solution 4: Excess deaths

In February 2021, half a year on since its August 2020 update, GOV.UK, while maintaining its official reporting policies and distinguishing itself from the ONS, considered yet another approach:

(5) But there is a third measure, which arguably provides the most comprehensive overview of the impact of the pandemic: excess deaths.

[35] As reported by GOV.UK: "Our review considered epidemiological evidence to see how likely it was that COVID-19 was a contributory factor to a death at different points in time after a positive test. […] Counting all deaths in people who have laboratory-confirmed infection […] is technically robust because it does not require a judgement to be made about cause of death." https://publichealthmatters.blog.gov.uk/2020/08/12/behind-the-headlines-counting-covid-19-deaths/

> These are the number of deaths over and above what would be expected, based on trends in previous years. Because they capture deaths from all causes – not just COVID-19 – they give us an idea of both the direct and indirect impact of the pandemic.[36]

So defined, the concept of EXCESS DEATHS has been gaining prominence in the discussions as the pandemic progressed and its impacts have become ever more apparent. Apart from *direct* COVID-19 deaths (notwithstanding all the methodological challenges on how to account for them, especially in the case of comorbidities such as cancer, hypertension, or diabetes), there is a large category of *indirect* deaths that includes: a) people who died of other conditions that appeared or aggravated during the pandemic but were not properly treated because of lack of access to healthcare, whether actual (discontinued treatments, cancelled operations, no hospital beds available) or perceived (fear of going to hospitals and contracting the virus) and b) people who suffered depression and other mental health issues, possibly leading to suicides. (All the same, due to reduced mobility and limited transmission of other viruses, there was also a marked *decrease* in mortality due to, e.g., traffic accidents or seasonal influenza.)

Among the institutions that proposed to refocus attention on the concept of EXCESS DEATHS are the Institute for Health Metrics and Evaluation (IHME), an independent population health research center at the University of Washington and the Center for Global Development, a think tank in Washington, D.C., that prepared a report on excess deaths in India, one of the countries hardest hit by the pandemic and also widely suspected of inefficient reporting of COVID-related data.[37] These institutions brought up two key concerns: 1) comparison of excess deaths to the estimated total (direct) COVID deaths and, in turn, those to the data on COVID deaths as officially reported by various governments; 2) the import of the concept of excess deaths itself.

[36] https://publichealthmatters.blog.gov.uk/2021/02/08/counting-deaths-during-the-pandemic/

[37] https://www.cgdev.org/sites/default/files/three-new-estimates-indias-all-cause-excess-mortality-during-covid-19-pandemic.pdf

As for 1), the IHME reports the following:

> (6) Deaths that are directly due to COVID-19 are likely underreported in many locations, particularly in settings where COVID-19 testing is in short supply. Most excess mortality is likely misclassified COVID-19 deaths. An analysis by the Netherlands statistical agency suggested that all excess deaths in the Netherlands were directly due to COVID-19. In fact, their analysis actually suggested that direct COVID-19 deaths may be higher than estimated excess deaths because deaths due to some other causes have declined during the pandemic. [38]

Moreover, drawing from different data sources, IHME evaluated the "ratios of total COVID-19 deaths to reported COVID-19 deaths": in their global tally, Belgium is among the countries with the lowest distortion of officially "reported" deaths to actual "total" deaths.[39] This indicates the Belgian broad concept might, in the end, have the sought-after empirical adequacy.[40]

Regarding 2): These complex comparisons and estimates make clear that EXCESS DEATHS are not being proposed as a more precise, simpler, or more fruitful version of the concept of COVID-19 DEATHS. Instead, they are being proposed as a concept that can, as it were, cut the knot and supersede the concept of COVID-19 DEATHS altogether. The argument for this conceptual replacement, rather than for continuous refinement of the notion of COVID-19 DEATHS, runs as follows: What counts in the bigger scheme of things—global public health, global economy, etc.—is the overall impact of the pandemic on the world's

[38] https://www.healthdata.org/research-analysis/diseases-injuries/covid/estimation-total-and-excess-mortality-due-covid-19

[39] https://www.healthdata.org/research-analysis/diseases-injuries/covid/estimation-total-and-excess-mortality-due-covid-19

[40] At the same time, the CGD reports that in India "the death toll from the pandemic is likely to be an order of magnitude greater than the official count of 400,000", namely, one of around 4mln. See https://www.cgdev.org/sites/default/files/three-new-estimates-indias-all-cause-excess-mortality-during-covid-19-pandemic.pdf

population. And the concept of excess deaths allows to gauge this impact in a more robust, adequate, and methodologically neat way. Excess deaths thus mark a conceptual shift similar, indeed directly related to, the shift from PANDEMIC to SYNDEMIC, discussed above in Section 2 (examples 1 and 1a).

4. Discussion

Our analysis lets us develop two points. First, metalinguistic arguments over "COVID-19 deaths" are inextricably linked to substantive, scientific issues and, as we hypothesize, are partly determined by the imperfect character of our epistemic position on the subject. Second, they work in the service of broader practical arguments whereby scientific results are weighted against broader public policy values (e.g., a broader definition might justify more decisive containment measures).

4.1. Between scientific and institutional concepts

What is at stake in the broader debate we analyzed is a natural expectation of a simple correct answer to the question "Which (and how many) deaths are due to COVID-19?" Nonetheless, as we have shown, there has been no single, privileged, natural concept of COVID-19 DEATH and no simple answer to this question. Principled concerns, most centrally related to the notion of "cause of death" in complex medical situations, make a clear classification of cases problematic, even assuming ideal access to the relevant information. Worse still, we are far from ideal access to the relevant information, as practical concerns of limited capacity for widespread testing and thorough autopsies have shaped the pandemic since its onset. Despite such concerns, in light of urgent need for public health intervention, some concrete response is needed. As we argued, uncertainties marring the *scientific concept* of COVID-19 DEATH recommend *metalinguistic intervention* on the *institutional concept*, designed to provide a fitting response to the circumstances. Such institutional interventions have two important features, already adumbrated by Searle.[41] First, reasoned control over meanings is well possible: being authoritative *declarations*, official interventions on the meaning of 'COVID-19 death' belong to the recognized deontic powers of the institutions mandated, among other things, to pronounce on the meaning of disputed terms. Second, such interventions are not entirely divorced from the attempts to get at the truth of the matter. They are, after all, *representative* declarations, expected to track, as much as possible, the features of the natural concept in question.

Taken together, these two features have some notable consequences. Metalinguistic interventions, as we understand them, are never closed or definitive. While they are meant to resolve some initial indeterminacy, they can typically not avoid all relevant sources of vagueness and indeterminacy. Indeed, if enough problematic cases accumulate after a first intervention, further action may be justified—as evidenced in the British case, where the definition of the COVID-19 deaths has been altered as new data became available. Such dynamicity of conceptual work allows to further understand the fertile tensions and interactions between "the facts on the ground" (revealed, e.g., via more accurate tests and autopsies) and the metalinguistic work performed by the institutions.

4.2. Metalinguistic interventions as practical arguments

All the interventions we analyzed, starting from the WHO's definition reflecting primarily "interests of importance for public health", also reveal the heterogeneity of factors shaping MIs. As we have amply illustrated, the aim of following, however approximately, the truth of the matter is only one of the interesting themes and determinants of MIs. Other considerations alien to the question of descriptive accuracy clearly

[41] "[I]n certain institutional situations we not only ascertain the facts but we need an authority to lay down a decision as to what the facts are after the fact-finding procedure has been gone through. [...] Some institutions require representative claims to be issued with the force of declarations in order that the argument over the truth of the claim can come to an end somewhere and the next institutional steps which wait on the settling of the factual issue can proceed" (Searle, 1975, p. 360).

contribute to the forging of institutional concepts. These are primarily practical concerns of public health policies: in case of any epistemic doubt, apply classification most conducive to battling the disease from the public health perspective (e.g., precautionary principle, the lesser risk, etc.).

As such practical normative grounds take precedence, it is worth reconstructing many of the reasons behind the metalinguistic declarations as instances of practical argumentation (Lewiński, 2017, 2018, 2021). Practical argumentation starts from an action-question: What shall we do under current (unwelcome) circumstances to reach the desired goals? These goals embody our main values. In our case, these are chiefly related to international public health—i.e., the prevention of deaths and disease, and control of the pandemic—and explicitly formulated in terms of availability of fast, frugal, and easily comparable data instrumental in efficient coordination among countries. Yet, over and above such health concerns, confidence of citizens in the institutions of the state, preservation of a good international image of a country, outlooks of economic recovery, etc., are also carefully balanced in addressing the practical question of which measures should be taken to best attain these heterogeneous goals. In the case of metalinguistic arguments, the measures to be taken, that is, the conclusion of a practical argument, is precisely the definitional declaration issued of the form: (*all things considered*, given our goals and values, under current circumstances and best knowledge we have,) we should count X as Y (see esp. Sec. 3.2).

As we have discussed earlier, many forms of metalinguistic arguments have been identified in the literature: arguments from analogy and from authority (Ludlow, 2014); dissociative arguments which split the current concept into two new concepts via a subscript gambit, e.g., COVID-19 $\text{DEATH}_{\text{SCIENTIFIC}}$ and COVID-19 $\text{DEATH}_{\text{INSTITUTIONAL}}$ (Chalmers, 2011; Pruś, 2021; Schiappa, 2003); as well as the whole wealth of definitional and semantic arguments, such as arguments from verbal classification (Pruś, 2021). On our analysis it seems, however, that the class of metalinguistic arguments is just coextensive with the class of arguments at large, in the sense of recognized forms of informal arguments. Concepts can be carved out and defended by analogy, authority, dissociation, example, causal relations, etc. In the context of our analysis, practical arguments to a specific definition have been particularly prominent.

All these forms of argumentation are surely worth investigating in terms of the role they play in metalinguistic interventions. Indeed, attention to argumentation lets us better see such interventions, which might otherwise remain inconspicuous even as they shape our collective lives. It also lets us better evaluate them: public arguments in support of such metalinguistic interventions should be explicitly made and open to scrutiny as publicly accountable forms of normative argumentation. With our analysis, we hope to have contributed to such scrutiny, however modestly.

Acknowledgements

This work has been supported by the COST Action project CA17132 and by the Portuguese Foundation for Science and Technology (FCT) and CHIST-ERA Programme via the project CHIST-ERA/0002/2019 – http://doi.org/10.54499/CHIST-ERA/0002/2019. We thank COST Action CA17132, European Network for Argumentation and Public Policy Analysis (http://publicpolicyargument.eu) for covering the Open Access fees of the original publication.

Sources

WHO:
https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020
https://www.who.int/docs/default-source/classification/icd/covid-19/guidelines-cause-of-death-covid-19-20200420-en.pdf

Belgium:
https://www.politico.eu/article/why-is-belgiums-death-toll-so-high/
https://www.politico.eu/article/in-defense-of-belgium-coronavirus-covid19-pandemic-response/
https://www.nytimes.com/2020/08/08/world/europe/coronavirus-nursing-homes-elderly.html
https://www.theguardian.com/world/2020/apr/24/is-comparing-covid-19-death-rates-across-europe-helpful-

UK:
https://blog.ons.gov.uk/2020/03/31/counting-deaths-involving-the-coronavirus-covid-19/
https://publichealthmatters.blog.gov.uk/2020/08/12/behind-the-headlines-counting-covid-19-deaths/
https://publichealthmatters.blog.gov.uk/2021/02/08/counting-deaths-during-the-pandemic/

IHME
https://www.healthdata.org/research-analysis/diseases-injuries/covid/estimation-total-and-excess-mortality-due-covid-19

CGD
https://www.cgdev.org/sites/default/files/three-new-estimates-indias-all-cause-excess-mortality-during-covid-19-pandemic.pdf

References

Allott, N., & Textor, M. (2012). Lexical pragmatic adjustment and the nature of *ad hoc* concepts. *International Review of Pragmatics, 4*(2), 185–208.

Amoretti, M. C., & Lalumera, E. (2021). COVID-19 as the underlying cause of death: Disentangling facts and values. *History and Philosophy of the Life Sciences*, *43*(1), 4. https://doi.org/10.1007/s40656-020-00355-6.

Antiochou, K., & Psillos, S. (2022). How to Handle Reasonable Scientific Disagreement: The Case of COVID-19. In S. Oswald, M. Lewiński, S. Greco, & S. Villata (Eds.), *The Pandemic of Argumentation* (pp. 65–83). Springer International Publishing. https://doi.org/10.1007/978-3-030-91017-4_4

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

Brun, G. (2016). Explication as a method of conceptual re-engineering. *Erkenntnis, 81*(6), 1211–1241.

Burgess, A., Cappelen, H., & Plunkett, D. (Eds.) (2020). *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Burgess, A., & Plunkett, D. (2013). Conceptual ethics I. *Philosophy Compass, 8*(12), 1091–1101.

Cappelen, H. (2018). *Fixing Language: An Essay on Conceptual Engineering.* Oxford: Oxford University Press.

Carnap, R. (1950). *The Logical Foundations of Probability*. Chicago: University of Chicago Press.

Chalmers, D. (2011). Verbal disputes. *Philosophical Review, 120*(4), 515–566.

Clarke, B., & Russo, F. (2016). Causation in medicine. In J.A. Marcum (Ed.), *The Bloomsbury Companion to Contemporary Philosophy of Medicine* (pp. 297-322). London: Bloomsbury.

Dutilh Novaes, C. (2020). Carnapian explication and ameliorative analysis: A systematic comparison. *Synthese, 197*(3), 1011–1034.

Dutilh Novaes, C. (2021). Argument and argumentation. In E.N. Zalta (ed.), *Stanford Encyclopedia of Philosophy* (Fall 2021 Edition). Available online: https://plato.stanford.edu/entries/argument/.

Eemeren, F. H. van, & Grootendorst, R. (1984). *Speech acts in argumentative discussions*. Dordrecht: Floris.

Eemeren, F. H. van, & Grootendorst, R. (2004). *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge: Cambridge University Press.

Eklund, M. J. (2021). Conceptual engineering in philosophy. In J. Khoo & R. Sterken (Eds.), *The Routledge Handbook of Social and Political Philosophy of Language* (pp. 15-30). New York: Routledge.

Elezkurtaj, S., Greuel, S., Ihlow, J., Michaelis, E. G., Bischoff, P., Kunze, C. A., Sinn, B. V., Gerhold, M., Hauptmann, K., Ingold-Heppner, B., Miller, F., Herbst, H., Corman, V. M., Martin, H., Radbruch, H., Heppner, F. L., & Horst, D. (2021). Causes of death and comorbidities in hospitalized patients with COVID-19. *Scientific Reports*, *11*(1), 4263. https://doi.org/10.1038/s41598-021-82862-5

Fairclough, I. (2022). The UK Government's "Balancing Act" in the Pandemic: Rational Decision-Making from an Argumentative Perspective. In S. Oswald, M. Lewiński, S. Greco, & S. Villata (Eds.), *The Pandemic of Argumentation* (pp. 225–246). Springer International Publishing. https://doi.org/10.1007/978-3-030-91017-4_12

Godden, D., & Casey, J. (2020). No place for compromise: Resisting the shift to negotiation. *Argumentation, 34*(4), 499–535.

Greco, S., & De Cock, B. (2021). Argumentative misalignments in the controversy surrounding fashion sustainability. *Journal of Pragmatics, 174*(1), 55-67.

Hall A. (2017). Lexical pragmatics, explicature and ad hoc concepts. In L. Depraetere & R. Salkie (Eds.), *Semantics and Pragmatics: Drawing a Line* (pp. 85-100). Cham: Springer.

Haslanger, S. (2012). *Resisting reality: Social construction and social critique*. Oxford: Oxford University Press.

Horn, L. R. (1985). Metalinguistic negation and pragmatic ambiguity. *Language*, *61*(1), 121-174.

Horton, R. (2020). Offline: COVID-19 is not a pandemic. *The Lancet*, 396(10255), 874. https://doi.org/10.1016/S0140-6736(20)32000-6

Ivanova, M. (2017). Aesthetic values in science. *Philosophy Compass, 12*(10), 1-9.

Johnson, R. (2000). *Manifest rationality*. Mahwah, NJ: Lawrence Erlbaum.

Koch, S. (2021). The externalist challenge to conceptual engineering. *Synthese*, *198*(1), 327–348.

Lindahl, B. I. B. (1988). On weighting causes of death. An analysis of purposes and criteria of selection. In A. Brändström & L.-G. Tedebrand (Eds.), *Society, health and population during the demographic transition* (pp. 131–156). Stockholm: Almqvist and Wiksell International.

Lindahl, B. I. B. (2021). COVID-19 and the selection problem in national cause-of-death statistics. *History and Philosophy of the Life Sciences*, *43*(2), 72. https://doi.org/10.1007/s40656-021-00420-8

Lewiński, M. (2017). Practical argumentation as reasoned advocacy. *Informal Logic*, *37*(2), 85-113.

Lewiński, M. (2018). Practical argumentation in the making: Discursive construction of reasons for action. In S. Oswald, T. Herman & J. Jacquin (Eds.), *Argumentation and language. Linguistic, cognitive and discursive explorations* (pp. 219-241). Cham: Springer.

Lewiński, M. (2021). Conclusions of practical argument: A speech act analysis. *Organon F, 28*(2), 420–457.

Ludlow, P. (2014). *Living words: Meaning underdetermination and the dynamic lexicon*. Oxford: Oxford University Press.

Machery, E. (2009). *Doing without concepts*. Oxford: Oxford University Press.

Marques, T. (2017). What metalinguistic negotiations can't do. *Phenomenology and Mind, 12*, 40–48.

Marques, T., & Wikforss, A. (2020*). Shifting concepts: The philosophy and psychology of conceptual variability*. Oxford: Oxford University Press.

Nilsson, L., Andersson, C., & Sjödahl, R. (2021). COVID-19 as the sole cause of death is uncommon in frail home healthcare individuals: A population-based study. *BMC Geriatrics*, *21*(1), 262. https://doi.org/10.1186/s12877-021-02176-z

Plunkett, D. (2015). Which concepts should we use? Metalinguistic negotiations and the methodology of philosophy. *Inquiry, 58*(7–8), 828–874.

Plunkett, D., & Sundell, T. (2013). Disagreement and the semantics of normative and evaluative terms. *Philosophers' Imprint, 13*(23), 1–37.

Plunkett, D., & Sundell, T. (2021). Metalinguistic negotiation and speaker error. *Inquiry*, *64*(1-2), 142-167.

Pruś, J. (2021). How can modifications of meaning influence argumentation? The concept and typology of semantic arguments. *Argumentation*, *35*(3), 483-508.

Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science, 7,* 131–193.

Rast, E. (2020). The theory theory of metalinguistic disputes. *Mind & Language*, DOI:10.1111/mila.12355.

Reiss, J. (2016). Causality and causal inference in medicine. In M. Solomon, J. R. Simon, & H. Kincaid (Eds.), *The Routledge Companion to Philosophy of Medicine* (pp. 58-70). New York: Routledge.

Reiss, J., & Ankeny, R. A. (2016). Philosophy of Medicine. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2016). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2016/entries/medicine/

Riggs, J. (2019). Conceptual engineers shouldn't worry about semantic externalism. *Inquiry*, *0*(0), 1–22. https://doi.org/10.1080/0020174X.2019.1675534

Rigotti, E., & Greco, S. (2019). *Inference in argumentation*. Dordrecht: Springer.

Sawyer, S. (2018). The importance of concepts. *Proceedings of the Aristotelian Society*, *118*(2), 127–147. https://doi.org/10.1093/arisoc/aoy008

Sawyer, S. (2020). Talk and thought. In A. Burgess, H. Cappelen, & D. Plunkett, D. (Eds.), *Conceptual engineering and conceptual ethics* (pp. 379-395). Oxford: Oxford University Press.

Scharp, K. (2020). Philosophy as the study of defective concepts. In A. Burgess, H. Cappelen, & D. Plunkett, D. (Eds.), *Conceptual engineering and conceptual ethics* (pp. 396-416). Oxford: Oxford University Press.

Schiappa, E. (1993). Arguing about definitions. *Argumentation, 7*(4), 403-417.

Schiappa, E. (2003). *Defining reality: Definitions and the politics of meaning*. Carbondale, IL: Southern Illinois University Press.

Searle, J. R. (1975). A taxonomy of illocutionary acts. In K. Günderson (Ed.), *Language, mind, and knowledge*, *vol. 7* (pp. 344–369). Minneapolis: University of Minnesota Press.

Searle, J. R. (1995). *The construction of social reality*. New York: Free Press.

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

Slater, T. A., Straw, S., Drozd, M., Kamalathasan, S., Cowley, A., & Witte, K. K. (2020). Dying 'due to' or 'with' COVID-19: A cause of death analysis in hospitalised patients. *Clinical Medicine*, *20*(5), e189–e190. https://doi.org/10.7861/clinmed.2020-0440

Soria Ruiz, A., (2021). Evaluative and metalinguistic disputes. *Australasian Journal of Philosophy*, online first: https://doi.org/10.1080/00048402.2021.1959624.

Sterken, R. K. (2020). Linguistic intervention and transformative communicative disruptions. In A. Burgess, H. Cappelen, & D. Plunkett, D. (Eds.), *Conceptual engineering and conceptual ethics* (pp. 417-434). Oxford: Oxford University Press.

Stojanovic, I. (2012). Emotional disagreement: The role of semantic content in the expression of, and disagreement over, emotional values. *Dialogue, 51*(1), 99-117.

Wilson, D. (2003). Relevance and lexical pragmatics. *Italian Journal of Linguistics*, *15*(2), 273-292.

The Argumentative Potential of Doubt: From Legitimate Concerns to Conspiracy Theories About COVID-19 Vaccines

Dima Mohammed and Maria Grazia Rossi

Mohammed, D. & Rossi, M. G. (2022). The argumentative potential of doubt: From legitimate concerns to conspiracy theories about COVID-19 vaccines. **In S. Oswald, M. Lewiński, S. Greco, & S. Villata (Eds.) *The Pandemic of Argumentation*. Springer. https://doi.org/10.1007/978-3-030-91017-4_7**

> If a man, holding a belief which he was taught in childhood or persuaded of afterwards, keeps down and pushes away any doubts which arise about it in his mind, purposely avoids the reading of books and the company of men that call into question or discuss it, and regards as impious those questions which cannot easily be asked without disturbing it– the life of that man is one long sin against mankind. (Clifford 1877, p. 5).

1. Introduction

In a recent study of the spread of anti-vaccine information on Facebook, researchers from George Washington University leave us with a distressing warning: the data we have today predicts that by the end of the decade, anti-vax viewpoints will become predominant (Johnson et al., 2020). The prediction is based on an analysis of the map of contention surrounding vaccines on the popular social platform. The map reveals highly dynamic interconnected clusters of anti-vaxxers, highly entangled with undecided clusters, while pro-vaccines clusters are rather peripheral. Beyond the alarming prediction, the study mirrors an equally alarming reality characterised by an explosive growth in anti-vaccination views and movements. While in principle, antivaccination views may be part of a healthy public debate about vaccines and public health, the growing spread of anti-vax emerges in the context of the proliferation of conspiracy theories sustained by a propagation of misinformation. As philosopher Lynch (2016) best captures it, the use of social media to spread misinformation is a "*giant shell game*": a golden deception opportunity for propagandists. As he argues, the danger of the increasing spread of misinformation is not just that it might lead people to believe in falsehood. While that is surely disturbing, what is equally perilous is that even if you are saved from false beliefs, misinformation can at least "get you confused enough so that you don't know what is true" (ibid). It is this power of 'manufacturing doubt' (Oreskes & Conway, 2010), which disinformation exercises even on the critical mind, that is most dangerous.

Indeed, doubt is a double-edge sword. On the one hand, uncertainty is an essential component in epistemic progress, and yet, doubt can also make us vulnerable to deception, confused to the point of no longer knowing what is true. Consider the difference between a scientist who designs a new experiment in order to verify an alternative hypothesis he suspects might be at play, and a politician who argues that energy policies do not need to change as long as there is still doubt that fossil fuel consumption is responsible for climate change. Or compare a patient's doubt about the efficacy of vaccines in stopping transmission expressed in a medical consultation, or an epidemiologist's suspicion about a potential link between a vaccine and blood clots expressed in a scientific meeting, to the doubt about the efficacy of vaccines expressed by a medical expert in a media interview. While in some cases it is rather clear that what is at stake is an expression of doubt that is benign or even epistemologically beneficiary, in other cases, the doubt seems to be rather tricky or even a typical example of a doubt manufactured in the context of a conspiracy theory.[1]

In this chapter, we explore doubt, its role, and the way it is being handled in the context of the public controversy about the COVID-19 vaccine. We examine anti-vaccine conspiracy theories from an argumentative perspective and analyse the argumentative potential that doubt can have in this public health controversy. Our analysis shows the importance of distinguishing between the different argumentative potentials a certain doubt can have. That, we argue, is necessary for an adequate response to the growing spread of conspiracy theories.

[1] As a precautionary note, we would like to make it clear that the discussion of the public controversy about the COVID-19 vaccine is not intended to establish the validity of medical facts. Despite the importance of such endeavour, our focus in this chapter is rather on the functioning of doubts typical of the COVID-19 vaccine controversy in the context of conspiracy theory.

2. Conspiracy Theories and the Argumentative Potential of Doubt

Broadly understood, a conspiracy theory (hereafter CT) is an alleged explanation of significant social and political events as the outcome of secret plots by two or more powerful actors (Aaronovitch, 2010; Byford, 2011; Coady, 2006; Dentith & Orr, 2018; Keeley, 2019). Following Oswald (2016, pp. 3–4), we examine conspiracy theories as argumentative objects: as communicative events that are embedded in controversy and disagreement, which intend to persuade a public of the conspiracy explanation by offering arguments in the premise-conclusion articulation. As argumentative objects, CTs have a common "argumentative profile": they make use of "source-related fallacies (...), hasty generalisations, arguments from analogy, inductive and abductive arguments, ad ignorantiam, and shifts in the burden of proof" (Oswald, 2016, p. 14). Furthermore, post hoc ergo propter hoc arguments are also very common, with anecdotal evidence and false correlations presented as scientific facts (Stolle et al., 2020). Argumentatively, CTs are "*refutational* narratives" (Byford, 2011) constructed in opposition to an official account of events rather than in justification of the conspiratorial account proposed (Oswald, 2016; Oswald & Herman, 2016; Wood & Douglas, 2013). Typically, CT's refutation is not much more than "the rhetoric of just asking questions" (Byford, 2011, pp. 88–93). Proponents of CTs pose questions to cast doubt on the official story (hereafter OS), focusing overly on data which the OS cannot account for and interpret the absence of answers as a cover up, a conspiracy to hide the truth (ibid; see also Stolle et al., 2020). Following Oswald's characterisation of the argumentative profile of CTs, in this chapter, we reconstruct the generic structure of a conspiracy theory macro argument. The reconstruction is based on a qualitative meta-analysis of conspiracy theories. In it, we propose a generic structure of the reasoning that links the different premises and argument types identified in the literature on CTs (e.g. Byford, 2011; Hofstadter, 1964; Jolley & Douglas, 2014; Lewandowsky et al., 2013; Nisbet, 2009; Oswald, 2016; Stolle et al., 2020; Zagarella & Annoni, 2019).

At the macro level, the explanation of social and political events as the outcome of secret plots by two or more powerful actors (Aaronovitch, 2010; Byford, 2011; Coady, 2006; Dentith & Orr, 2018; Keeley, 2019) may be considered the ultimate conclusion of any given CT. As such, CTs can be characterised as discourses advancing the claim that a certain official story x is the sinister work of powerful individuals and groups 'conspiring' against the general public. Challenging the official account (Oswald, 2016; Oswald & Herman, 2016; Wood & Douglas, 2013) is the main argument advanced in support of such a claim. Obviously, the justificatory power of this argument is problematic in a way that reflects a central problem of CTs. At best, the justificatory power at work is an argumentum ad ignorantiam: even if indeed the OS at stake were not credible, it would be just a too "big leap from the undeniable to the unbelievable" (Hofstadter, 1964, p. 35) to conclude that this is evidence for a conspiracy. The "big leap", which we take to be a central element of CTs, turns the CT argument inherently fallacious. In supporting the argument that the OS is not credible, proponents of CTs present evidence (real and fake) that goes against the OS and attack the credibility of the sources – the supposed political and social elites which includes authorities and experts representing the OS. Mistrust of official sources has indeed been a crucial element in the success of any conspiracy theory (Jolley & Douglas, 2014; Lewandowsky et al., 2013; Nisbet, 2009; Oswald, 2016).

Unlike the main standpoint, (1) *An official story x is the sinister work of powerful individuals and groups 'conspiring' against the general public*, which is often left implicit, the main premise (1.1) *The official story is not credible* is often expressed explicitly. Nevertheless, the great bulk of CT explicit discourse supports premises 1.1.1 and 1.1.2. In arguing that *The 'official sources' of OS x cannot be trusted* (1.1.1), CT rely on 'source-related arguments' (Oswald, 2016), typical examples allege that the proponents of OS x have vested interests (among other types of ad hominem arguments). In supporting that *There is evidence against what the official story says* (1.1.2), CT advocates present examples (real and fake) that contradict the OS. Interestingly, the more examples we have to support 1.1.2, the more 1.1.1 is supported too. In other words, 1.1.2 supports 1.1.1 too.

While we reconstruct the argument underlying CTs, it is crucial to keep in mind that in any given CT, there is no single homogeneous unified argument made explicitly by a single CT proponent. Instead, conspiracy theories are made up of various argument lines, articulated more or less explicitly by different individuals and groups. The individuals and groups may be in conflict one with the other and may vary in the degree of doubt they cast on the official story, from moderate scepticism all the way to denialism (Capstick & Pidgeon, 2014; Dunlap, 2013; Grimes, 2021; Haltinner & Sarathchandra, 2021; Pierre, 2020). Nevertheless, the diverse contributions converge into a discourse that defends a conspiracy explanation of a certain significant event. The reconstruction we propose is meant as a generic structure that represents exactly that: the CT argument as a discourse – an argument that is made up by the various contributions of different arguers advanced at different occasions. As we propose such a structure, we make no claims about the intentions of groups and individuals that contribute to the CT discourse. Not every arguer who expresses a certain CT premise is necessarily intending to convey the conclusion of the CT argument. Nevertheless, even without that intention on behalf of the arguer, the premise would still contribute to the CT discourse by invoking the conclusions associated with it. It is important to distinguish the intention of the arguer from the contribution the argument can make. Of course, both are important, and obviously the two can overlap, but the argumentative potential an argument has is not restricted to the justificatory force intended by the arguer. Distinguishing between the two is important in order to account for the way public arguments work without over-attributing commitments to arguers.

Generally speaking, the argumentative potential refers to the possible argumentative inferences a certain discursive choice can activate beyond what is explicitly stated. Think of the affirmation “my body, my rule”. A common argumentative potential associated with the statement is opposing the control and criminalization of sexuality and reproduction. The affirmation has been associated with the defense of the position in such a way that the two have formed a premise-conclusion pair, an inference, that is publicly recognisable. Whenever the affirmation is made, the position is invoked, even if it is not explicitly articulated. One way of capturing the argumentative potential is to identify premise-conclusion pairs that have become publicly recognizable, and in the absence of evidence to the opposite, affirming (x) may be interpreted as also claiming (y), on the basis that x has become publicly associated with the justification of y (Mohammed, 2019a).[2] The starting point here is an understanding that public arguments do not start from void, nor do they happen in isolation: every time an argument is made, it builds on already existing (lines of) arguments in which some premise-conclusion pairs become recognisable. While arguers may not be held committed to the argumentative potential of their premises beyond doubt, the commitment is rather presumptive (ibid.), in the discourse, premises have the potential of invoking the conclusions which are typically associated with them. The point here is not making claims about the intention of the arguer, but rather about the interpretation of the argument. This is crucial in public arguments, where what matters is not just what meaning a speaker intends to convey, but also what meaning is conveyed, on the basis of the already recognised premise-conclusion pairs and independent of the intention of the arguer.[3]

That discursive choices acquire argumentative potentials beyond what is explicitly said is in line with the idea that there is an argumentative aspect inherent in every form of language (Anscombre & Ducrot, 1983) as well as with the understanding that intertextuality and interdiscursivity are two fundamental aspects of discourse (Reisigl & Wodak, 2016;

[2] The most basic argumentative potential might be found in enthymemes where the conclusion is unexpressed. But the argumentative potential is not necessarily always as obvious nor necessarily intended as the implicit conclusion of a typical enthymeme is. See Mohammed (2019b) for more on this.

[3] The activation of an unexpressed inference might be achieved by a certain choice of proposition, as well as by the word choice and formulations used in the propositions. A skilled arguer would carefully make her discursive choices in order to convey intended messages as well as to avoid conveying unintended ones, i.e. to activate desired argumentative potentials as well as to curb undesired ones (Mohammed, 2019a, 2019b). Paying attention to the argumentative potential of discursive choices is crucial for the analysis and evaluation of arguments, especially arguments about socio-political issues made publicly. It is beneficial in order to capture the strategic shape of arguments (Mohammed, 2019a), as well as to explain how public misunderstandings arise and polarisation in public controversies deepens (Mohammed, 2019b).

Wodak, 2009).[4] Indeed, in today's networked public sphere (Benkler, 2006; Kaiser et al., 2018; Pfister, 2014), the argumentative potential is hardly ever confined to a single text or even a discourse: at any point in time, there are countless interrelated controversies being fed with new premises and conclusions as well as by the new inferences that connect them. Arguments emerge to manage the disagreement (Jackson & Jacobs, 1980; Jacobs & Jackson, 1989) as part of a complex network where distinct lines in relation to different issues crisscross and overlap (Aakhus,2002; Lewinski & Mohammed, 2015; Mohammed, 2019b). In such a complex network, where the boundaries are fluid and dynamic, the argumentative potential proliferates making it a tricky task to curb undesired potentials and activate only desired ones.[5]

In the next sections, we will examine the argumentative potential of doubt in the public arguments about COVID-19 vaccine. In particular, we examine how doubt functions in the context of conspiracy theories. We examine CT discourse through the generic argumentative structure sketched above. The structure allows us to see how the different parts of CT discourse hang together, to highlight what is common between the different CTs and to explain how they are interrelated, which is crucial for examining the argumentative potential of doubt. For example, the structure allows us to show how it is that "evidence for one conspiracy theory becomes evidence for all of them" (Byford, 2011); it shows how easily it is for a premise that discredits an 'official source' in a new CT to become just another piece of evidence for mistrusting the Official Story in general. Finally, as the analysis we conduct in the next sections will show, the reconstruction of the generic CT argument allows us to shed light on the manufacturing of doubt typical of CT discourse.

[4] Furthermore, Reisigl and Wodak (2016) consider that that discourse is characterised by (a) macrotopic-relatedness, (b) pluri-perspectivity related to various voices in a specific social field, and (c) argumentativity.

[5] See Mohammed (2019b) for na example of the complexity of managing the argumentative potential in a public controversy.

3. COVID-19 Vaccine: The Conspiracy Theory

Conspiracy theories about the COVID-19 pandemic emerged as soon as the pandemic itself became a global reality (Ellis, 2020). In these theories, which have been typically accompanied by disinformation campaigns, one may identify a few common themes (Grimes, 2021, pp. 3–4). The most general of these themes is the claim that COVID-19 is an outright hoax, or alternatively that it has been deliberately engineered, in both cases in order to suppress freedoms on a global scale.[6] Other main conspiratorial themes advance that COVID-19 is a pretext for a mass vaccination programme in which philanthropist Bill Gates is going to microchip people to spy on them and eventually control them, or that the pandemic has been caused by 5G electromagnetic radiations (ibid.). These and other themes have been circulated widely by people from all walks of life including by "leaders and people in positions of trust and authority" (Douglas, 2021, p. 272). The role celebrities and public figures play in creating and feeding CTs cannot be exaggerated, especially considering social media. In late March 2021, a study by the Center for Countering Digital Hate (CCDH) and Anti-Vax Watch revealed that up to two thirds of anti-vaccine content circulating on major social media networking sites can be traced back to 12 individuals and their organizations. The twelve anti-vaxxers have since then been dubbed the "Disinformation Dozen" (CCDH, 2021).[7]

Vaccine conspiracy theories are by no means a new phenomenon. Since the first claims were made in the 1990s about a link between the MMR vaccine and autism, the anti-vax movement has never disappeared. It was only to be expected that as soon as talk of COVID-19 vaccine began, a new conspiracy theory emerged. Looking at the history of the modern anti-vax movement, Stolle et

[6] Interestingly, "While such narratives seem mutually opposed, they are frequently held in tandem by a cohort of believers despite mutual exclusivity – a not infrequent situation with conspiratorial thinking" (Grimes, 2021, p. 3).

[7] The "Disinformation Dozen" is made up of Ty and Charlene Bollinger, Robert F. Kennedy Jr., Joseph Mercola, Sherri Tenpenny, Rizza Islam, Rashid Buttar, Erin Elizabeth, Sayer Ji, Kelly Brogan, Christiane Northrup, Ben Tapper, and Kevin Jenkins.

al. (2020) identify common argumentative patterns of anti-vaccination proponents. Medical mistrust and other forms of anti-system arguments (e.g., medicine as a profit-making enterprise); fear of adverse consequences, caused in the case of the MMR by the association with autism spectrum disorder, as well as of other neurological disorders, and finally fear of harmful ingredients contained by vaccines (ibid.). Many of the premises remained very similar when the COVID-19 vaccine CT emerged. In particular fears of side effects, and the chronic mistrust in medical authorities (Rief, 2021; Verger & Dubé, 2020). Just like other CTs, the COVID-19 vaccine conspiracy is characterised by central tenets which are reasonably consistent, and yet which manifest themselves in a diversity of narratives, worldviews and ideologies, and express varying degrees of doubt about the official story. From the libertarian gun rights advocates in the US, to leftist big pharma sceptics in France and anti-lockdown activists both on the far left and the far right in Germany, anti-vaccine conspiracy theories allege that we have been lied to about the pandemic: about its origin, magnitude but most importantly, about the vaccine story we are being told. The different anti-vaccine conspiracy theory narratives converge, without necessarily agreeing on the nature of the conspiracy, nor on the extent to which the conspiracy is the work of a sinister powerful elite that works against the general public. Furthermore, the COVID-19 vaccine CT is intertwined with other COVID-19 CTs (e.g., lockdown, masks … etc.) as well as other CTs in general (e.g., QAnon). As the analysis in the next section will show, this openness is an important power house for conspiracy theories.

In order to discredit the official story about the COVID-19 vaccine, conspiracy theories manufacture doubt in relation to five main areas. First, doubt is raised about the safety of the vaccine as a cornerstone of the vaccine OS: Is the vaccine really safe or does it cause serious dangerous side-effects? Doubt about vaccine safety is raised by focusing on the occurrence of side effects as well as by alleging that the clinical trials to produce the COVID-19 vaccine have been rushed in a way that compromises its safety. Second, doubt is raised about the effectiveness of the vaccine: Is the vaccine really as effective in combatting the pandemic as it is claimed to be? Third, vaccine CT questions the threat of COVID-19 as alleged by the official story: Is COVID-19 as dangerous as it is being presented by medical authorities and experts? Fourthly, doubt is also raised in relation to the composition of the vaccine: is the vaccine ethically produced or does it contain harmful substances? This doubt links the vaccine CT to the QAnon CT which alleges that vaccines are bioweapons developed by elite paedophile networks. Finally, doubt is raised about trust in the official medical experts and authorities, the proponents of the official story: can we really trust the profit-making big pharma enterprises? Can we trust the medical authorities, for example in view of their history of unethical treatments of minorities and people of colour? Or yet more generally, can we trust that the 'system' is really trying to save us? Here too, the overlap with other CTs such as QAnon is obvious.

In what follows, we look into each of these lines of doubt. We spell out their argumentative role in the CT and give examples of the instantiations in its discourse, particularly in the discourse of the Disinformation Dozen.

(a) Is the vaccine really safe as claimed?

In the COVID-19 vaccine CT, doubt about vaccine safety is manufactured to discredit the OS by supporting the CT premise 1.1.2 (in Fig. 1), namely that there is evidence against what the COVID-19 OS says. Anecdotal accounts of people dying after they get vaccinated are the most common examples. Here is one, presented by Robert F. Kennedy, Jr., the head of the Children's Health Defense and probably the most visible and vocal member of the Disinformation Dozen. It is a piece of news that appears under the Big pharma news section on Kennedy's organisation's page. The news reads as follows:

(1) 58-Year-Old Woman Dies Hours After Getting First Dose of Pfizer Vaccine. Doctors said Drene Keyes, whose death is under investigation, died of flash pulmonary edema likely caused by anaphylaxis, a life-threatening allergic reaction, which some people have experienced after receiving the COVID vaccine (Children's Health Defense, 2021)

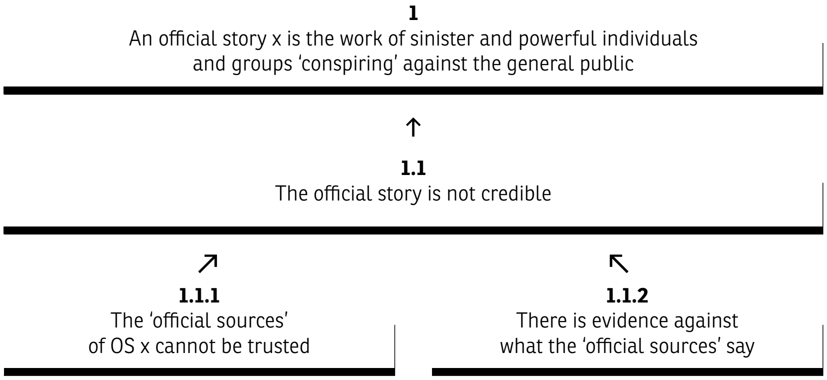

1
An official story x is the work of sinister and powerful individuals and groups 'conspiring' against the general public
1.1
The official story is not credible
1.1.1
The 'official sources' of OS x cannot be trusted
1.1.2
There is evidence against what the 'official sources' say

FIGURE 1
Generic structure of a conspiracy theory argument

Leaving aside the factual accuracy of the news, it is interesting that the case, which is presented as evidence that the vaccine can kill you, can also cast doubt on what the OS says. The news has a clear potential of feeding mistrust in the official medical institution as well.

(b) Is the vaccine really as effective as claimed?

Anti-vaccine CTs employ the doubt about vaccine effectiveness as another line of evidence against what the COVID-19 OS says (CT premise 1.1.2 in Fig. 7.1). This is a line of argument that has been pursued by Joseph Mercola, the American alternative medicine proponent and co-author of the book *The Truth About COVID-19* (Mercola & Cummins, 2021). In the book, the authors do not understate their claims:

(2) Effectiveness of the vaccines has been wildly exaggerated and major safety questions have gone unanswered (Chelsea Green Publishing, 2021).

Here too, the formulation of the affirmation activates not just the argumentative potential to undermine the accuracy of the OS, but also that of undermining the trust in the official sources.

(c) Is the COVID-19 disease really the threat it is presented to be?

The seriousness of the COVID-19 disease is at the core of the OS about the pandemic. Therefore, raising doubt about it has the obvious argumentative potential of undermining the OS (CT premise 1.1.2 in Fig. 7.1). Interestingly, many national medical groups have also been expressing this doubt. For example, in a video shared at the World Doctors Alliance, Dutch general practitioner Elke De Klerk says:

(3) We do not have a pandemic. COVID-19 is a normal flu virus (Newswise, 2020).

This is one of the doubts most propagated by public figures, starting with the Brazilian president Jair Bolsonaro who spoke of COVID-19 as a "little flu" but not ending with Donald Trump who has in September 2020 retweeted a message claiming "the true number of COVID-19 deaths in the United States was a small fraction of the official numbers".

(d) Is the vaccine ethically produced?

In the context of CT, alleging information that casts doubt on the production of the vaccine fulfils the argumentative potential of lending direct support to the premise that *The proponents of the COVID-19 OS cannot be trusted* (CT premise 1.1.1 in Fig. 7.1). Consider the following example. Reporting on an interview with obstetrics and gynaecology physician Christiane Northrup, another one of the Disinformation Dozen, the NOQ Report website (Scheuer, 2020) tells us that:

(4) Dr. Northrup discussed the questionable composition of the vaccines being readied, and noted that they likely include fetal materials coming from babies aborted in China, as well as other materials that allow the tracking of individuals and their movements. Oddly, it seems that China sent the disease to the United States, and now it is making a profit from supplying materials from aborted babies for the coming vaccines.

Interestingly enough, NOQ Report is a news and opinion website that states as its mission the fighting of "fake news by the mainstream media" (Scheuer, 2020). It is simply in line with the website's "mission" to manufacture doubt in order to foster the conspiratorial potential associated with discrediting the sources associated with the OS.

(e) Are the official medical experts and authorities worthy of public trust?

Casting doubt on the trustworthiness of official sources, experts and medical authorities is one of the most powerful doubts manufactured by CTs. Undermining the trust in the official sources does not just play directly into discrediting the OS. It also lends support to the ultimate CT claim that the OS is the work of a group conspiring against the general public. It is therefore not surprising that this doubt is often expressed in combination with other doubts, such as in examples (i), (ii) and (iii) above. In the context of the COVID-19 vaccine, two paths to undermine trust have been popular: big pharma purportedly using immunization as a mere profit-making enterprise, and medical authorities accused of continuing a history of unethical treatments of minorities and people of colour. Of the Disinformation Dozen, social media influencer Rizza Islam has an Instagram account dedicated to fueling mistrust in the medical authorities. In his "Not Another Tuskegee Experiment", the African American activist invokes the legacy of the abusive Tuskegee Study[8] to feed an already existing trust problem. The argumentative potential is rather clear, and yet, it surely does not harm the CT to repeat it. In a Facebook live-stream, Kevin Jenkins (CEO of the Urban Global Health Alliance and another member of the Disinformation Dozen) spoke to the Black community about the COVID-19 vaccine:

(5) They are spending a trillion dollars to convince you that it's OK to kill yourselves (McGill Office for Science & Society, 2021).

Undermining an already shaky trust in medical authorities is a fast track towards supporting the ultimate claim of the COVID-19 vaccine CT. Moreover, it is an unignorable contribution in support of other CTs that would flourish every time trust is undermined in another official story.
Having seen how the different manufactured doubts typical of the

[8] The Tuskegee Study of Untreated Syphilis in the Negro Male was conducted between 1932 and 1972 by the United States Public Health Service and the Centers for Disease Control and Prevention. The purpose of the study was to observe the natural history of untreated syphilis. To achieve it, black men with syphilis were left untreated to essentially see what would happen (Brawley, 1998; Centers for Disease Control and Prevention, 2021).

COVID-19 vaccine controversy are employed in the context of conspiracy theory, in the next section, we will look more into how doubt about the vaccine functions beyond the CT discourse. We will analyse doubt about vaccine safety in general and discuss different argumentative potentials of this doubt. The discussion will explain how the same doubt can be considered a legitimate expression of ambivalence but may also be used as part of a more articulated sceptic position, or even as evidence for a conspiracy theory that casts doubt on an official story altogether.

4. Handling the Argumentative Potential: Doubt About the Safety of COVID-19 Vaccine

In general, doubt about vaccine safety is one of the main doubts expressed when considering the COVID-19 vaccine. Concerns about safety arose as early as talk about the vaccine began, especially given the speed in which COVID-19 vaccines were developed and approved compared to previous vaccines. As news was reporting the progress in developing the new vaccines, the public was being reminded that "The vaccine development process has typically taken a decade or longer" (Thompson, 2020). The impression was created that in order to respond to the urgency of developing a vaccine, the clinical trials phase was cut short which might eventually compromise the certainty about vaccine side-effects. The doubt about safety, in particular, concerns about serious side-effects, grew as the trials got repeatedly halted because of suspicions about side-effects. Eventually the trials resumed, and vaccines were approved. Nevertheless, doubt about safety re-emerged and grew yet stronger as a result of the repeated news about the occurrence of blood clots post vaccination, as well as the recurrent halt in administering the vaccine by the medical authorities (Wise, 2021).

Doubt about safety of vaccines is in principle legitimate. In general, this doubt is an integral part of the development of any vaccine: it is the doubt that underlies the clinical trials, and which guides the precautionary halt in both trials and the roll-out once there is suspicion that a certain vaccine is causing an unforeseen side-effect. In its lightest manifestation, the doubt is a form of incertitude about the possibility of side-effects that can compromise the safety of the new vaccine: clinical trials are designed for scientists to rule this doubt out and present a convincing case in support of the vaccine's safety. Yet, the doubt about side-effects can be stronger, for example, as it happened with the *Oxford-AstraZeneca* vaccine, it can be motivated by a repeated occurrence of blood clots post vaccination, or by a recurrent halt in administering the vaccine by the medical authorities (Wise, 2021). Although these may be legitimate reasons to cast doubt on the safety of a vaccine, such motivated doubt needs to be handled carefully (Wadman, 2020). Unless the *argumentative potential* of doubt is controlled, it is a slippery slope where doubt can slither quickly from natural ambivalence to legitimate scepticism all the way to conspiracy theory denialism. As the analysis below will show, what distinguishes between these three are the different argumentative potentials that can be associated with the reason motivating the doubt.

Let us take the example of the doubt about the safety of the Oxford-AstraZeneca vaccine motivated by the fact that several people have died from unusual blood clots after getting the vaccine (EMA, 2021, April 7). The reason motivating the doubt, namely that *several people have died from unusual blood clots after getting the Oxford-AstraZeneca vaccine*, has at least three argumentative potentials:

1. Ambivalence

Considering that *several people have died from unusual blood clots after getting the Oxford-AstraZeneca vaccine* may give rise to the minimum degree of doubt about the safety of the vaccine: ambivalence on whether or not the vaccine is safe, without necessarily leaning to any of the positions. From an argumentative perspective, expressing ambivalent doubt amounts to assuming the dialectical role of the antagonist in a non-mixed dispute (Van Eemeren & Grootendorst, 1992, pp. 16–22) concerning the standpoint challenged by the reason motivating the doubt. The position may be reconstructed as: *several people have died from unusual blood clots after getting the Oxford-AstraZeneca* therefore *I am not sure if the vaccine is safe or not*. In this case, what underlies the ambivalence is uncertainty about the causal link between the vaccine and the reported blood clots. In other words, even though the doubt is motivated by the possibility of such a link, the link itself is subject of doubt.[9] Ambivalent doubt is the type of doubt that gave rise to the precautionary measures taken by medical authorities in countries that halted vaccine roll-out until the causal link is investigated and doubt about the safety is ruled out.

On its bearer, ambivalent motivated doubt incurs no obligation apart from the willingness to give up the doubt if the reasons motivating the doubt get adequately addressed. On the proponents of the position challenged, the obligation is obviously higher: medical authorities, as well as the pharma, are expected to adequately respond to the ambivalent doubt by addressing its motivating reasons. Ambivalence is the minimum argumentative potential that a motivated doubt can have. It can be that it is all there is at stake in an argumentative situation, but more often than not, motivated doubt can activate higher argumentative potentials.

2. Scepticism

In addition to ambivalence about whether or not the Oxford-AstraZeneca vaccine is safe, the fact that several people have died from unusual blood clots after getting the vaccine can give rise to vaccine safety scepticism. Assuming that there is a causal link between the vaccine and the unusual blood clots, the motivated doubt acquires the potential to function as an argument against the position that the vaccine is safe. The position of sceptic doubt may be reconstructed as: *several people have died from unusual blood clots after getting the vaccine therefore I do not think that the vaccine is safe*. In argumentative terms, this amounts to assuming the dialectical role of the protagonist in a mixed dispute about the safety of the vaccine. A sceptic position about the safety of the Oxford-AstraZeneca vaccine incurs on its bearer an obligation that mirrors the obligation of the opponents of the vaccine safety thesis. Medical authorities and pharma ought to justify why the vaccine may still be considered safe despite the

[9] A relevant factor here might also be related to the definition of drug safety in general. Even if it is accepted that there is a causal link between the harm observed and the drug, how much risk is tolerated before a certain drug is no longer considered safe? Ambivalence can be the result of uncertainty about that, and misunderstanding can result from a mismatch about the definition of drug safety between communicators.

unusual blood clots, and vaccine safety sceptics ought to defend that in view of the unusual blood clots the vaccine may not be considered safe. Scepticism is a medium range argumentative potential when it comes to doubt about vaccine safety. Scepticism goes further than ambivalence in that it assumes a position concerning vaccine safety while ambivalence does not, but just like in ambivalence, the argumentative potential of a sceptic doubt remains within the dispute over vaccine safety. While that is surely possible, doubt about vaccine safety may also have argumentative potential that extends beyond that dispute.

3. Denialism

An important far-reaching argumentative potential of the vaccine safety doubt is the one associated with anti-vax CT movements. As we have seen in the previous section, doubt about vaccine safety makes an important line in the vaccine conspiracy theory argument. Conspiracy theorists take advantage of every new case of serious vaccine side-effects, presenting it as yet another evidence against the *official story* which alleges that the vaccine is safe. Interpreted within the conspiracy theory argument, the doubt motivated by the occurrence of unusual blood clots can acquire the following CT denialist potential:

Several people have died from unusual blood clots after getting the Oxford-AstraZeneca,

This is (yet another) evidence that the vaccine is not safe,

Therefore, the official story about the vaccine is not credible.

The doubt motivated by possible serious side-effects has been used in its denialist potential over and over by vaccine conspiracy theories, i.e. as an argument to discredit the *official story* about vaccines altogether. What we have here is an inference, a premise-conclusion pair, which has become publicly recognisable: new evidence that the vaccine is not safe is a sign that the official story about the vaccine is not credible. The conclusion, namely that *the official story about the vaccine is not* *credible*, is hanging out there as a standing standpoint (Mohammed, 2019a) waiting for the premise to be expressed so that it may take effect. The denialist argumentative potential functions by virtue of this public inference, that is by virtue of the premise conclusion pair being recognised and invokable. Whenever there is a new reason motivating the doubt about the vaccine safety, there is an argumentative potential for the doubt to take the denialist direction. Furthermore, another publicly recognizable inference at work here is the one that leads to the main CT claim: *The OS about the vaccine is not credible therefore The COVID-19 vaccine official story is the work of sinister and powerful individuals and groups 'conspiring' against the general public*. In both cases, the potential is there; whether it materialises or not depends on the way arguers interpret each other's arguments.

Obviously, the denialist potential is problematic. To start with, it is based on a flawed inference. At best, it is a hasty generalisation to discredit the official vaccine story altogether even if it were true that the vaccine is not safe (which in itself is the conclusion of another hasty generalisation). But that is not all. In the discourse of conspiracy theorists, flawed reasoning is typically combined with the spread of misinformation. False accounts of vaccine-related deaths as well as exaggerations of side-effects reports are circulated to sustain the false generalisation, which leads to growing levels of vaccine hesitancy, one of the main public health challenges in the context of the current COVID-19 pandemic (Pullan & Dey, 2021; Weintraub et al., 2021; World Health Organization, 2020).

Furthermore, what may be even more problematic than the flawed reasoning underlying the denialist argumentative potential is the way that potential can distort positions and unnecessarily polarise the public discussion. It is indeed a tricky task to know which argumentative potential is most adequate when an arguer expresses a motivated doubt. It is not always easy to know whether a speaker who reports that *Several people have died from unusual blood clots after getting the Oxford AstraZeneca* is expressing ambivalence on whether the vaccine is safe or not, or if she is being rather sceptic that the vaccine is safe, or if she is even presenting the news as evidence that we cannot trust official authorities and their vaccine claims. Misunderstandings can happen if an arguer and their interlocutor interpret the doubt in terms of different argumentative potentials. Ideally, a competent arguer should be capable of curbing an argumentative potential that is undesired to her. The simplest way to do that is using a disclaimer: for example, an arguer who is aware that their ambivalence might be misunderstood as scepticism might choose to explicitly affirm that they are "not saying that the vaccine is not safe".[10] Nevertheless, in public controversies, arguers may not be always aware of a certain argumentative potential that can be ascribed to them, which eventually complicates the task of controlling how they are being interpreted (see examples in Mohammed, 2019a, 2019b). Furthermore, the task is even more difficult in a polarised context, characterised by conspiracy theories. The louder the conspiracy theories, the more present their public inferences are, and the more likely it is that the denialist argumentative potential is wrongly attributed to expressions of motivated doubt that are meant in non-denialist potentials. Indeed, in the public discussion about the COVID-19 vaccine, doubt has too often been misinterpreted as an expression of the denialist stance leaving people feeling misinterpreted and alienated (Douglas et al., 2019; Stolle et al., 2020).

In spite of the difficulty of identifying the argumentative potential at stake, medical experts and authorities, proponents of the vaccine safety thesis in general, are under the obligation of responding to doubt about their thesis. Ambivalent and sceptic doubt can disappear if evidence is provided. In response to the doubts motivated by post-vaccine blood clots, an effective answer has been provided by comparing the risk of blood clots post-vaccination with that associated with other medication considered safe. For example, experts explained that the risk of clots with the Oxford AstraZeneca vaccine is roughly 1:250,000, while the risk of clots for the contraceptive pill is 1:2000 (Mahase, 2021). The comparison would probably not remove a denialist doubt, but it is quite likely that it is effective in overcoming cases of ambivalent and even sceptic doubt. While non-denialist doubts can be overcome, doubts ignored are prone to getting hijacked by conspiracy theories who transform the neglect into yet another reason to discredit the official story and its proponents. The official sources do not respond because they do not have an adequate answer, or because they do not even care, goes the typical conspiracy theory.

[10] Obviously, such a disclaimer might be interpreted as a case of a rhetorical apophasis. The arguers might watch out for that for it can backfire.

5. Discussion

How to respond to conspiracy theories is undoubtedly a pressing urgent question. For as Douglas (2021,p.271) puts it, "conspiracy theories are consequential, and in many studies have been linked to climate denial, vaccine refusal, political apathy, apathy in the workplace, prejudice, crime, and violence". Various strategies for addressing the consequences of CT have been suggested in the literature. One strategy has been confrontation. For example, Romer and Jamieson (2020, p. 113355) argue that "Because belief in COVID-related conspiracy theories predicts resistance to both preventive behaviours and future vaccination for the virus, it will be critical to confront both conspiracy theories and vaccination misinformation to prevent further spread of the virus in the US." Romer and Jamieson recommend "continued messaging by public health authorities on mainstream media and in particular on politically conservative outlets that have supported COVID-related conspiracy theories" (ibid.). In the same vein, Douglas (2021, p. 272) suggests that "'inoculating' people with factual information can stem the influence of conspiracy theories". However, confronting the conspiracy is a risky choice. The allure of conspiracist explanation lies to a great extent in their simplification, rather oversimplification, of complex realities. It might be overly optimistic to believe that the rather more complex truth would simply win the public's mind once they are presented with it. Just consider how little success it has yielded to fact-check the misinformation presented as part of the different CTs in the last decades. Furthermore, explicitly engaging with conspiracy theories risks giving them more presence.

There is a danger that the more we engage with CTs, the more publicly present conspiracist inferences become, and the harder it gets to avoid interpreting uncertainty in a denialist argumentative potential. But while engaging with conspiracy theories is surely not the answer, ignoring them is not either. It might be understandably tempting to think that the right thing to do is to ignore, or even delegitimize the doubts that fuel conspiracy theories. Indeed, that has been the predominant attitude when it comes to vaccine-related CTs. The history of the never-ending

MMR vaccine controversy is a good example (see Jackson, 2020). But conspiracy theories fuel on doubts, and ignored doubts do not disappear. To the contrary, ignoring them is turned in itself into another piece of evidence in favour of the conspiracy. What is needed is an approach that addresses the doubts hijacked by CTs without giving presence to the CTs themselves. That would be an approach that engages with doubt, but not with its denialist argumentative potential.

There is indeed a need to reconsider the ease in which doubt is being interpreted as an expression of a conspiracy theory, for as it signals irrationality, the CT label can neutralize valid concerns and delegitimize people (Douglas et al., 2019; Harambam & Aupers, 2017; McKenzie-McHarg & Fredheim, 2017; Orr & Husting, 2018; Räikkä & Basham, 2018). But reconsidering the CT label only begins by acknowledging the legitimacy of doubt, and it is not completed until different argumentative potentials are assigned to the different types of doubt. Distinguishing between different argumentative potentials is a crucial element in a response that acknowledges legitimate concerns without empowering conspiracy theories. It is in a sense a way to avoid that an unnecessarily broad interpretation of conspiracy theory dominates the public debate and leaves an uncertain public a prey to it. It is important to distinguish between different argumentative potentials but when that is not possible, medical authorities should interpret doubt in the ambivalent potential. Ambivalent doubt ought to be addressed by experts and health authorities who have the adequate knowledge to respond to the reason motivating it.

A final word, on the argumentative potential of doubt in its relation to trust. Indeed, CTs cannot be countered without addressing the question of trust. In order to reduce the impact of conspiracy theories, Nisbet (2009) suggests that “trusted messengers” are employed. As she explains, combating the conspiracy theory may be likely to have more success if the counterarguments come from trusted sources such as valued ingroup members, instead of outgroup members who are typically associated with mistrust (ibid.). The “trusted messengers” strategy seems to have been guiding Dr. Anthony Fauci, Director of the US National Institute of Allergy and Infectious Diseases, as he fostered partnership with African American groups and religious leaders. Also in the same vein, it has been a news highlight that Moderna’s COVID-19 vaccine is being studied by a team of scientists led by a black woman, Dr. Kizzmekia Corbett. While it is surely helpful to present the public with sources they trust, an adequate response ought to also curb the argumentative potential that doubt can have in undermining trust. In CT discourse, doubt is presented as evidence against the OS. But that can be successful only if doubt is not already part of the OS. In other words, the argumentative potential of doubt to discredit the OS might disappear if doubt is integrated in the OS. While ambivalent doubt is surely already part of the vaccine OS, more communicative effort is needed to present it as such: to present an OS that is more realistic and therefore not easily discredited by doubt.

References

Aakhus, M. (2002). Modeling reconstruction in groupware technology. In F. H. van Eemeren (Ed.), *Advances in pragma-dialectics* (pp. 121–136). SicSat.

Aaronovitch, D. (2010). *Voodoo histories: The role of the conspiracy theory in shaping modern history*. Riverhead Books.

Anscombre, J. C., & Ducrot, O. (1983). *L'argumentation dans la langue*. Pierre Mardaga.

Benkler, Y. (2006). The wealth of networks: How social production transforms markets and freedom. In *The wealth of networks: How social production transforms markets and freedom*. Yale University Press. https://doi.org/10.2307/20455766

Brawley, O. W. (1998). The study of untreated syphilis in the negro male. *International Journal of Radiation Oncology, Biology, Physics, 40*(1), 5–8. https://doi.org/10.1016/s0360-3016(97)008 35-3

Byford, J. (2011). *Conspiracy theories: A critical introduction*. Palgrave Macmillan.

Capstick, S. B., & Pidgeon, N. F. (2014). What is climate change scepticism? Examination of the concept using a mixed methods study of the UK public. *Global Environmental Change, 24*(1), 389–401. https://doi.org/10.1016/j.gloenvcha.2013.08.012

CCDH. (2021). *The disinformation dozen: Why platforms must act on twelve leading online antivaxxers*. Retrieved from https://www.counterhate.com/disinformationdozen

Centers for Disease Control and Prevention. (2021). *The Tuskegee timeline*. Retrieved from https:// www.cdc.gov/tuskegee/timeline.htm

Chelsea Green Publishing. (2021). *The truth about COVID-19*. Retrieved from https://www.chelse agreen.com/product/the-truth-about-covid-19/

Children's Health Defense. (2021). *58-year-old woman dies hours after getting first dose of pfizer vaccine*. Retrieved from https://childrenshealthdefense.org/defender/woman-dies-hoursafter-first-dose-pfizer-vaccine/

Clifford, W. K. (1877). The ethics of belief (1877) I. The duty of inquiry. *Contemporary Review*.

Coady, D. (2006). *Conspiracy theories: The philosophical debate*. Ashgate.

Dentith, M. R. X., & Orr, M. (2018). Secrecy and Conspiracy. *Episteme, 15*(4), 433–450. https:// doi.org/10.1017/epi.2017.9

Douglas, K. M. (2021). COVID-19 conspiracy theories. *Group Processes and Intergroup Relations, 24*(2), 270–275. https://doi.org/10.1177/1368430220982068

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology, 40*(S1), 3–35. https://doi.org/ 10.1111/pops.12568

Dunlap, R. E. (2013). Climate change skepticism and denial: An introduction. *American Behavioral Scientist, 57*(6), 691–698. https://doi.org/10.1177/0002764213477097

Ellis, E.G.(2020). The coronavirus outbreak is a petri dish for conspiracy theories. *Wired*. Retrieved from https://www.wired.com/story/coronavirus-conspiracy-theories/

EMA. (2021). *COVID-19 Vaccine Janssen: Assessment of very rare cases of unusual blood clots with low platelets continues*. European Medicines Agency. Retrieved from https://www.ema.eur opa.eu/en/news/covid-19-vaccine-janssen-assessment-very-rare-cases-unusual-blood-clots-lowplatelets-continues

Grimes, D. R. (2021). Medical disinformation and the unviable nature of COVID-19 conspiracy theories. *PLoS ONE, 16*(3 March), 1–17. https://doi.org/10.1371/journal.pone.0245900

Haltinner, K., & Sarathchandra, D. (2021). Considering attitudinal uncertainty in the climate change skepticism continuum. *Global Environmental Change, 68*, 102243. https://doi.org/10.1016/j.glo envcha.2021.102243

Harambam, J., & Aupers, S. (2017). 'I am not a conspiracy theorist': Relational identifications in the dutch conspiracy Milieu. *Cultural Sociology, 11*(1), 113–129. https://doi.org/10.1177/174997 5516661959

Hofstadter, R. (1964). *The paranoid style in American politics and other essays*. Harvard University Press.

Jackson, S. (2020). Evidence in Health Controversies. In *OSSA Conference Archive, 15*. Retrieved from https://scholar.uwindsor.ca/ossaarchive/OSSA12/Friday/15

Jackson, S., & Jacobs, S. (1980). Structure of conversational argument: Pragmatic bases for the enthymeme. *Quarterly Journal of Speech, 66*(3), 251–265. https://doi.org/10.1080/003356380 09383524

Jacobs, S., & Jackson, S. (1989). Building a model of conversational argument. In B. Dervin, L. Grossberg, B. J. O'Keefe, & E. Wartella (Eds.), *Rethinking communication: paradigm exemplars* (Vol. 2, pp. 153–171). Sage.

Johnson, N. F., Velásquez, N., Restrepo, N. J., Leahy, R., Gabriel, N., El Oud, S., & Lupu, Y. (2020). The online competition between pro- and anti-vaccination views. *Nature, 582*(7811), 230–233. https://doi.org/10.1038/s41586-020-2281-1

Jolley, D., & Douglas, K. M. (2014). The effects of anti-vaccine conspiracy theories on vaccination intentions. *PLoS ONE, 9*(2). https://doi.org/10.1371/journal.pone.0089177

Kaiser, J., Fähnrich, B., Rhomberg, M., & Filzmaier, P. (2018). What happened to the public sphere? The networked public sphere and public opinion formation. *Handbook of Cyber Development, Cyber-Democracy, and Cyber-Defense*, 433–459. https://doi.org/10.1007/978-3319-09069-6_31

Keeley, B. L. (2019). Of conspiracy theories. *Conspiracy Theories: The Philosophical Debate, 96*, 45–60. https://doi.org/10.4324/9781315259574-4

Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2013). NASA faked the moon landing-therefore, (Climate) science is a Hoax: An anatomy of the motivated rejection of science. *Psychological Science, 24*(5), 622–633. https://doi.org/10.1177/0956797612457686

Lewinski, M., & Mohammed, D. (2015). Tweeting the Arab Spring: Argumentative polylogues in digital media. In C. Palczewski (Ed.), *Disturbing argument* (pp. 291–297). Routledge.

Lynch, M. P. (2016, November 28). Fake news and the internet shell game. *The New York Times*. Retrieved from https://www.nytimes.com/2016/11/28/opinion/fake-news-and-the-internet-shellgame.html

Mahase, E. (2021). AstraZeneca vaccine: Blood clots are "extremely rare" and benefits outweigh risks, regulators conclude. *Bmj*, (April), n931. https://doi.org/10.1136/bmj.n931

McGill Office for Science and Society. (2021). *A dozen misguided influencers spread most of the anti-vaccination content on social media*. Retrieved from https://www.mcgill.ca/oss/art icle/covid-19-health/dozen-misguided-influencers-spread-most-anti-vaccination-content-socialmedia

McKenzie-McHarg, A., & Fredheim, R. (2017). Cock-ups and slap-downs: A quantitative analysis ofconspiracyrhetoricintheBritishParliament1916–20151. *HistoricalMethods, 50*(3),156–169. https://doi.org/10.1080/01615440.2017.1320616

Mercola, J., & Cummins, R. (2021). *The truth about COVID-19: Exposing the great reset, lockdowns, vaccine passports, and the new normal*. Chelsea Green Publishing.

Mohammed, D. (2019a). Managing argumentative potential in the networked public sphere: The anti- # MeToo Manifesto as a case in point. In B. Garssen, D. Godden, G. R. Mitchell, & J. H. M. Wagemans (Eds.), *Proceedings of the 9th conference of the international society for the study of argumentation* (pp. 813–822). Sic Sat.

Mohammed, D. (2019). Standing standpoints and argumentative associates: What is at stake in a public political argument? *Argumentation, 33*(3), 307–322. https://doi.org/10.1007/s10503-0189473-y

Nisbet, M. C. (2009). Communicating climate change: Why frames matter for public engagement. *Environment: Science and Policy for Sustainable Development, 51*(2), 12–23. https://doi.org/10. 3200/ENVT.51.2.12-23

Newswise. (2020). *A video posted by a European-based group called World Doctors Alliance falsely claims the novel coronavirus is "a normal flu virus"*. Retrieved from https://www.newswise.com/factcheck/a-video-posted-by-a-european-based-group-calledworld-doctors-alliance-falsely-claims-the-novel-coronavirus-is-a-normal-flu-virus

Oreskes, N., & Conway, E. M. (2010). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Press.

Orr, M., & Husting, G. (2018). Media marginalization of racial minorities: “conspiracy theorists” in U.S. Ghettos and on the “Arab street.” In J. E. Uscinski (Ed.), *Conspiracy theories and the people who believe them* (pp. 82–93). Oxford University Press. https://doi.org/10.1093/oso/978 0190844073.003.0005

Oswald, S. (2016). Commitment attribution and the reconstruction of arguments. In F. Paglieri, L. Bonelli, & S. Felletti (Eds.), *The psychology of argument* (pp. 17–32). College Publications.

Oswald, S., & Herman, T. (2016). Argumentation, conspiracy and the moon: A rhetorical-pragmatic analysis. In M. Danesi & S. Greco (Eds.), *Case studies in discourse analysis* (pp. 295–300). Lincom Europa.

Pfister, D. S. (2014). *Networked media, networked Rhetorics – Attention and deliberation in the early blogosphere*. The Pennsylvania State University Press.

Pierre, J. M. (2020). Mistrust and misinformation: A two-component, socio-epistemic model of belief in conspiracy theories. *Journal of Social and Political Psychology, 8*(2), 617–641. https:// doi.org/10.5964/jspp. v8i2.1362

Pullan, S., & Dey, M. (2021). Vaccine hesitancy and anti-vaccination in the time of COVID-19: A Google Trends analysis. *Vaccine, 39*(14), 1877–1881. https://doi.org/10.1016/j.vaccine.2021. 03.019

Räikkä, J., & Basham, L. (2018). Conspiracy theory phobia. In J. E. Uscinski (Ed.), *Conspiracy theories and the people who believe them* (pp. 178–186). Oxford University Press. https://doi. org/10.1093/ oso/9780190844073.003.0011

Reisigl, M., & Wodak, R. (2016). The discourse-historical approach. In R. Wodak & M. Meyer (Eds.), *Methods of critical discourse analysis* (3rd ed., pp. 23–61). Sage.

Rief, W. (2021). Fear of adverse effects and COVID-19 vaccine hesitancy. *JAMA Health Forum, 2*(4), e210804. https://doi.org/10.1001/ jamahealthforum.2021.0804

Romer, D., & Jamieson, K. H. (2020). Conspiracy theories as barriers to controlling the spread of COVID-19 in the U.S. *Social Science and Medicine, 263*, 113356. https://doi.org/10.1016/j.soc scimed.2020.113356

Scheuer, M. (2020, December 1). *Dr. Christiane Northrup explains the science and politics behind COVID-19 and the coming vaccines. Are we being told the truth about COVID-19? NOQ Report*. Retrieved from https://noqreport.com/2020/12/01/dr-christiane-northrup-explains-the-scienceand-politics-behind-covid-19-and-the-coming-vaccines/?fbclid=IwAR3L7KBLaEa-B9GPgc bF4FpN1P9jbD9UUPwt3DyLUU_kJEGmiLLFVfxW93E

Stolle, L. B., Nalamasu, R., Pergolizzi, J. V., Varrassi, G., Magnusson, P., LeQuang, J. A., & Breve, F. (2020). Fact vs fallacy: The anti-vaccine discussion reloaded. *Advances in Therapy, 37*(11), 4481–4490. https:// doi.org/10.1007/s12325-020-01502-y

Thompson, S. A. (2020, April 30). How long will a vaccine really take? *The New York Times*. Retrieved from https://www.nytimes.com/interactive/2020/04/30/opinion/coronavirus-covid-vac cine.html

Van Eemeren, F., & Grootendorst, R. (1992). *Argumentation, communication and fallacies. Erlbaum.*

Verger,P. & Dubé, E. (2020). Restoring confidence in vaccines in the COVID-19 era. Expert *Review of Vaccines, 19*(11), 991–993. https://doi.org/10.1080/14760584.2020.1825945

Wadman, M. (2020). Public needs to prep for vaccine side effects. *Science, 370*(6520), 1022. https:// doi.org/10.1126/science.370.6520.1022

Weintraub, R. L., Subramanian, L., Karlage, A., Ahmad, I., & Rosenberg, J. (2021). Covid-19 vaccine to vaccination: Why leaders must invest in delivery strategies now. *Health Affairs, 40*(1), 33–41. https://doi.org/10.1377/hlthaff.2020.01523

Wise, J. (2021). Covid-19: How AstraZeneca lost the vaccine PR war. *The BMJ*, 373, n921. https:// doi.org/10.1136/bmj.n921

Wodak, R. (2009). *The discourse of politics in action: Politics as usual*. Palgrave Macmillan. https:// doi.org/10.1057/9780230316539

Wood, M. J., & Douglas, K. M. (2013). What about building 7? A social psychological study of online discussion of 9/11 conspiracy theories. *Frontiers in Psychology, 4*(Jul). https://doi.org/10. 3389/fpsyg.2013.00409

World Health Organization. (2020). *Improving vaccination demand and addressing hesitancy*. Geneva. Retrieved from http://awareness.who.int/immunization/programmes_systems/vac cine_hesitancy/en

Zagarella, R. M., & Annoni, M. (2019). A rhetorical perspective on conspiracies. *Journal of Argumentation in Context, 8*(2), 262–283. https://doi.org/10.1075/jaic.18006.zag

Metaphors and Persuasion in Healthcare Communication

Maria Grazia Rossi

Rossi, M. G. (2021). Metaphors and persuasion in healthcare communication.
Langages, 222(2), 59-76.
https://doi.org/10.3917/LANG.222.0059

1. Metaphorical persuasion through implicit arguments

This paper will analyse the epistemic and ethical function of metaphors to determine if and how they can promote a 'common ground' between patients and healthcare providers. We focus exclusively on metaphors that are related to health conditions/care (Sopory 2017). Specifically, we examine how metaphors can serve as educational tools to promote:

(1) patient understanding of a particular health condition, or how the patient should interpret medical evidence;
(2) shared decision making in choosing from among different therapeutic options.

Metaphors have a great pedagogical potential (Cameron 2003, Ortony 1975, Rossi 2016). As far back as Aristotle, scholars have described metaphors as a transference of meaning, *i.e.*, a way to understand something in terms of something else (Black 1955 and 1979, Gibbs 1994, Kövecses 2015, Lakoff & Johnson 1980). One such example is talking about *diabetes* (target) as a monster (source), where the chronic disease is reasoned in terms of a threatening and scary imaginary creature.[1] By explaining diabetes as a monster, a specific figurative frame is provided, organising information of what diabetes is and how it works.

Understanding diabetes through this metaphor is to adopt a persuasive strategy that directs attention to specific aspects of this health condition: the example of the monster metaphor frames diabetes by giving salience to specific characteristics of the target; that is, it stresses the importance of dominating diabetes to avoid future risky complications. From a theoretical point of view, as clearly stated by Entman (1993, 52):

[1] Visit the MySugr website to know more about how the monster metaphor is used to describe diabetes, people with diabetes, and how they feel about diabetes. In this article, all the examples that extend or specify this metaphor were found on the MySygr website: http://mysugr.com/diabetes-monster-friend-or-foe/ (accessed on December 10th, 2023).

> To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described.

In this regard, metaphors are effective persuasion and framing tools because as a result of their exposition, people can change their intentions, beliefs, and behavioural attitudes (Petty & Cacioppo 1986).

In the context of health discourse, persuasion through metaphorical frames can be exploited to influence citizens' intentions to adopt specific social policies and behaviours, improving engagement in health-focused behaviours (Scherer, Scherer & Fagerlin 2014). Metaphors can modify people's intentions, as indicated by research around health-related issues such as vaccination, anti-smoking ad campaigns, or cancer prevention behaviours (Harrington 2012, Hauser & Schwarz 2014, Scherer, Scherer & Fagerlin 2014). Moreover, patients commonly use metaphors in sharing their emotions and experience of their illness, which perhaps explains why the use of metaphors by healthcare providers is perceived positively by patients, both in terms of how patients assess the providers' communication skills and in terms of the greater trust they place in the healthcare relationship (Casarett *et al.* 2010, Demjén & Semino 2016, Gibbs & Franks 2002, Harrington 2012, Semino *et al.* 2015).

It has been argued that metaphors exploit implicit arguments in order to persuade, working as devices that promote reasoning and argumentation (Ervas, Gola & Rossi 2018, Rossi 2016, Macagno & Zavatta 2014; but see Doury 2003). This idea has been discussed in the context of argumentation theory mainly considering the analogical arguments behind the use of a metaphor. For instance, Perelman and Olbrechts-Tyteca (1969: 399) wrote: "In the context of argumentation, at least, we cannot better describe a metaphor than by conceiving it as a condensed analogy, resulting from the fusion of an element from the *phoros* with an element from the Theme". With respect to the present study, Perelman and Olbrechts-Tyteca's observation is useful in pointing out that metaphors always contain implicit or condensed arguments in support of a specific view of the target; these implicit arguments may encourage a certain line of reasoning, and coherently, also promote behaviours or (at least) behavioural intentions.

Implicit arguments behind metaphorical persuasion allow describing how metaphors can be used as educational tools that favour the building of common ground between interlocutors, thus contributing to shared understanding and decision-making. The efforts made by interlocutors to build common ground have been referred to in pragmatics generally, and intercultural pragmatics more specifically, as efforts to converge on a shared representation of the reality obtained by the activation of or seeking in memory of the shared knowledge, but also by creating new knowledge within the communicative process (Kecskes & Zhang, 2009). Metaphors are particularly suitable in this regard (Jacobi 1984). First, by proposing an interpretation of unfamiliar concepts (target) through more familiar ones (source), metaphors can be used to explain new concepts and gain knowledge. Second, in proposing to look at the target from a specific (but also partial) perspective, metaphors can be used to favour some inferential pathways over others. Namely, metaphors provide just a partial representation of the reality, since they foreground specific features of the target; exploiting the similarities with the source, they background other (sometimes relevant) features of the target (Burgers, Konijn & Steen 2016, Semino 2008).

A deeper analysis of the diabetes-monster metaphor can be useful to illustrate how reasoning may be influenced through implicit arguments, and it is also paradigmatic in emphasizing what is, at the same time, the greatest potential and weakness in using a metaphorical frame. Starting from what the monster metaphorical personification can explain, let us come back to the inferences that can be drawn from the metaphorical frame "diabetes is a monster": a person with diabetes (the monster) is represented as a monster tamer, so that when the diabetes monster is tamed (that is, well-managed), it may appear as a foe that deserves ball-and-chain with a zipped-up mouth.

In this respect, the monster metaphor seems to combine two arguments. In the first place, drawing on the dangerous nature of monsters, this metaphor implicitly appeals to the undesirable consequences diabetes may lead to in terms of future risky complications (e.g., cardiovascular disease, diabetic retinopathy, diabetic nephropathy, lower limb amputation). In the second place, this metaphor also suggests a solution. Representing the patient as a monster tamer, such a metaphor implicitly proposes a cause and effect reasoning, based on which it is possible to deal with the monster problem (and its negative consequences): if diabetes complications (effect) are due to a monster that is not tamed (cause), then it is possible to tackle the cause by taming the monster to avoid unpleasant effects. This interpretation sheds light on the potential of the monster metaphor: a brave and lifelong struggle (again, a metaphor) against the monster seems to be the main point of the arguments on which this metaphor is based.

However, it appears more difficult to extrapolate a more positive idea of what it might mean to live with diabetes from this metaphor as if, for example, the diabetes-monster were a travel companion or a friend (both metaphors used by patients and healthcare providers). The monster metaphor is not immediately conducive to thinking about a tamed monster that becomes a friend, and that is also able to celebrate the victory of the monster tamer. This is also how people with diabetes may feel when diabetes is going well.

On the one hand, the monster metaphor seems adequate in that it points out the importance of self-management in diabetes care, given that poorly compensated diabetes is risky and causes serious complications; but, on the other hand, this metaphor may not appear to be the best metaphor to show that living compliantly with diabetes leads to a happy life. It may be problematic and dangerous, impacting negatively on patients' engagement and wellbeing. Especially in the context of communicating risk as in healthcare, it is imperative to identify metaphors as more or less appropriate. This is even more important because the monster metaphor does exist, and it is indeed the main narrative standpoint around which the mySugr mobile application and diabetes advice revolves. In this regard, it is even more interesting that on the website of the mySugr company, users are asked to answer the following question "Your diabetes monster – friend or foe?" This request is obviously (and perhaps primarily) a way to know how people feel about their life with friend/foe diabetes. However, it also makes explicit the difficulty of accounting for the positive dimensions of good diabetes compensation within storytelling that is completely based on the monster metaphor.

In a post published in an Italian Facebook peer support group for persons with diabetes (see also Rossi & Menichetti 2019), one member wrote a long message describing an episode of anger and fragility caused by a hypoglycemic event occurring during the night.[2] The message is as follows:

> **Example 1**
> Ma davvero pensano ancora che il diabete sia una malattia da niente? È un essere che mi mangia piano piano e sono costretta a farmelo amico.
>
> Do they really think that diabetes is an easy disease? It is a being who eats at me slowly and I am being forced to make him my friend. (*translation mine*)

This example shows very well how the two metaphorical descriptions (diabetes as a foe *vs.* diabetes as a friend) contradict each other, framing different aspects of diabetes that are difficult to reconcile. However, in the case of the participant/patient, the concomitant use of these two metaphorical frames makes it possible to express (emotional) difficulties that healthcare providers (but also families and friends) must understand and try to alleviate, even before revealing if they have therapeutic recommendations to suggest. Vice versa, as in the case of the mySugr mobile application or in the hypothetical case of a provider using only the monster metaphor, the risk in relying on this partial frame is to cause unexpected side effects (such as discouraging patients from imagining a happy life despite diabetes).[3]

This article discusses examples of metaphorical frames present in data of medical consultations to identify the contextual appropriateness

[2] I thank the participant for giving me permission to use her words in this article.

[3] Note that the metaphorical comparison "diabetes is a friend/travel companion" is not without risk, as Example 1 shows very well. Perhaps it is not a coincidence that people with diabetes sometimes modify this metaphor as follows: "diabetes is an unwelcome friend/uninvited guest".

of metaphors and how persuasive strategies can be effectively used to improve understanding and decision-making in the context of patient-provider interactions. The relevance of metaphors as an educational tool to improve understanding is also considered within intercultural and cross-cultural medical contexts, setting priorities for future research.

2. Medical recommendations as acts of persuasion

The medical interview as an institutional context has been described as an advice-seeking discourse type, a complex interactional activity in which the recommendations received by the advisee are assumed to be the best viable option (Bigi 2018). In the words of Costello and Roberts (2001, 244) "a recommendation is nothing more than a proposal for a course of action". Patients have an active decision-making role and adopt various linguistic strategies to show that they agree with or refuse the recommendations; in other words, a medical recommendation is the product of a negotiation between patients and healthcare providers. On their part, healthcare providers must assess the medical situation, interpret the evidence, and consider their patients' preferences and values: that is, healthcare providers need to offer reasons in favour of a given medical recommendation. On these grounds, Feng and colleagues (Feng, Bell, Jerant & Kravitz 2011) referred to a recommendation as a genuine "act of persuasion", where physicians should even act as advocates to persuade patients (see also Barilan & Weintraub 2001, Stahl 1998).

The complexity of such a social and linguistic interaction has led scholars to question how a medical recommendation should be structured in order to be effective, which also means securing patient adherence to treatment. In this respect, Bigi (2018) highlighted the importance of argumentation sequences (with persuasion being one of the tools available within them), while Feng and colleagues (2011) investigated the role of persuasion strategies in different dimensions of a medical recommendation. To explain why an integrated model of medical advising is needed, Feng *et al.* (2011: 287) pose the following question:

> If we conceptualize medical recommendations as a form of persuasion, a question that must be posed is this: "What can physicians say and do to be an effective advocate of a treatment plan?"

This question can be interpreted as referring to two different levels of analysis: a linguistic level, taking into account the linguistic strategies favouring persuasion, and an ethical level, accounting for which of those linguistic-persuasive strategies can be deemed legitimate.

Regarding the linguistic level of analysis, Feng *et al.* (2011) chose to investigate the persuasive role of explicit communication strategies (seriousness of the problem, treatment effectiveness, patient's self-efficacy, and potential limitations with the recommended treatment), obtaining ambiguous results. Commenting on their findings, they wrote: "Simply addressing a dimension may not be sufficient to elicit corresponding perception from the patient; a certain degree of emphasis may be necessary" (*ibid.*, 2011: 294). However, implicit persuasion strategies such as metaphors may also play a role in this context. Metaphors could be used as implicit persuasive strategies to emphasize key issues in (medical) conversations. People cannot avoid using metaphors, especially when they are talking about complex issues, both when these issues concern a complexity related to the emotional and the informational/decisional side of the disease (Macagno & Rossi 2019). Metaphors appear to be more effective for understanding than their literal counterpart (Sopory 2017, Sopory & Dillard 2002, Van Stee 2018) when the latter is available. However, metaphors can sometimes be dangerous, causing misunderstanding within patient-provider interactions (Macagno & Rossi 2019; see also Section 3 below).

On the ethical level of analysis, persuasion should be conceived within an ethical framework (Kunneman *et al.* 2019; Rubinelli 2013) and, what is more relevant to the purpose of this paper, through an ethical use of metaphors. Indeed, metaphorical frames (and the implicit arguments of which they are made) may be improperly used, even against the interests of those who introduced them.

An example drawn from the context of Assisted Reproductive Technology (ART) will make clear the relevance of the ethical dimension behind the use of metaphors. Example 2 is an excerpt from a deliberative sequence coming from a doctor-couple consultation conducted in Italy (Borghi *et al.* 2018; Leoni *et al.* 2018). The ineffective communicative and argumentative quality of the extract has already been described (Rossi, Leone & Bigi 2017).[4] Little has been said on the ethical role played by metaphorical frames in defining the communicative and argumentative quality of the ART field.

The excerpt refers to the part of the consultation in which the Doctor (D) is helping the couple to complete the informed consent form when the Female Patient (FP) starts to explain why she wants to do just one single attempt at assisted reproduction (from line 4). The discussion around the metaphorical frame "undergoing ART treatments is/is not forcing nature" represents the key point of the whole discussion.

Example 2

The example has been translated into English; the original transcript is in Italian. Transcription conventions refer to the Jefferson Transcription System: [] square brackets indicate overlapping talk; = the equals sign indicates the end of one line and beginning of the next with no gap/pause in between (sometimes a slight overlap if there is a change of speaker); ::: colons indicate vowel or consonant lengthening. A complete table of the symbols used within the Jefferson Transcription System is available here: https://www.universitytranscriptions.co.uk/jefferson-transcription-system-a-guide-to-the-symbols/ (accessed on December 10th, 2023).

1 D since it's better to use a bigger number of egg cells, we can't freeze them, [otherwise]
2 FP [no::: no::: no no (unint)]
3 D so, no, we start all over again

[4] Example 2 was already discussed in Rossi, Leone and Bigi (2017) as an example of the improper use of argumentative instruments. Here the focus is on the improper use of a metaphorical frame by the healthcare provider. For a more detailed discussion of all the argumentative passages see the original paper.

4 FP no, I already decided to go for one try

5 FP and that’s it, because, I think, I mean, I don’t think I would be able be able to… start all over again another time. I mean, if it’s God’s will, otherwise it’s like starting a farm…

6 D wow, you sure sound negative, don’t you?

7 FP [I’m not being negative], I’m a little fatalist

8 FP because, I feel that I am already forcing a bit… what is supposed to be, [I mean…]

9 D [but why (unint)]?

10 FP ah, I don’t know, but… that’s it

11 MP[5] well, doc, she’s always been kind of negative about kids

12 FP yeah, I mean, it’s not like I’ve ever been head over heels about kids, I mean, it’s not like I’m dying to become a mother. I realize it’s something he really wishes, it’s probably the age. Kids are cute, all right, but when I was in my thirties I was thinking, no way, I don’t want any. Then you grow older and maybe you change your mind, maybe [the context]

13 D [things change]

14 FP things change a bit. But it’s not like I’ve always thought that I wanted to be a mother. No, I wanted to be a woman, a daughter, that’s it. So, I’ve already tried, did everything that was possible, treatm- everything, ‘cause, the past four years we’ve spent always travelling around the place…

15 FP this is the last time, I’m trying once and then [then that’s it]

16 D [listen]

17 FP [because I’m fata-]

18 MP [listen to me, doc, in the end]

19 FP [because] I’m fatalist

20 FP because then, I see people who don’t have any children, people who get children… what if you get a child…

[5] The Male Patient is labeled MP.

that's not one hundred per cent… I know myself, so

21 D yeah, well, all right, but then [in any case technology (unint)]

22 FP [I know that but then…] yeah, sure, techn- of course, but, you know, I'm already forcing the hand…. For me this is forcing nature

23 D we sure are funny, aren't we? (chuckling softly)

24 D you know why, I was thinking, we never have these thoughts

[look]

25 D for example, you get pneumonia

26 FP it's true

27 D and you take antibiotics, when you get cancer- now [mind you, I'm not putting them on the same level]

28 FP [yeah, of course not, no no no]

29 D but it's funny though, because then you don't think that you're forcing nature, and instead on this thing about children

30 D [do you know why] I'm telling you? Because it's something I get from so many [couples]

31 FP [really?] eh

32 D it's something a lot of people feel, this thing about forcing nature because probably it really comes=

33 MP [and then after all]-

34 D [=it's felt] like something that [should be natural]

35 FP [should probably be natural] it's all, mm… a cultural thing we carry with us, I don't know if it's something…

36 D I guess so

37 FP yeah, probably it's all a cultural thing, not anything else

38 D that is rooted

39 FP that is rooted in- in-… all that catholic thing and bla bla bla you grow up with, it's probably that, but then in the end it's such a part of you that=

40 FP = for me, that I didn't even want to become a mother, when I was… I mean, we started late for that reason,

because when I was thirty the last thing I wanted was
to become a mom so… now I'm forty and at this point
I think, if I make it that's good, otherwise I go on too
much and I feel like a grandma and I don't… I mean, I
get all those thoughts, that when my child is thirty I'm
seventy [all this kind of stuff, you know, so]
41 FP one thing- one time, I try
42 MP Sure
43 FP and then
44 D ok, so, this decision is very [personal]
45 FP [sure]
46 D and I really don't want to interfere because…
47 FP no no
48 D although I would really like to tell you something, that
will maybe make it a little easier for you

FP introduces some ideas and arguments to motivate her conviction (see in particular lines 12, 14, and 20), including the use of the metaphorical frame mentioned above (lines 8, 22), which close FP's argument (*I'm already forcing the hand…. For me this is forcing nature*, at line 22). In the current context, FP's metaphorical argument "ART treatment is forcing nature" should be interpreted in light of what FP has already shared with D, which is strongly emotionally connotated, not least because of her life plan (*But it's not like I've always thought that I wanted to be a mother. No, I wanted to be a woman, a daughter, that's it*, at line 14).

The argumentative strategy advanced by D latches onto the metaphorical frame of FP on other inferential pathways, which do not seem to be those favoured by FP. More particularly, D calls into question the cogency of the metaphorical frame used by FP (lines 29, 32, 34), extending the frame to other disease situations in which medical treatments are used without giving the impression that we are forcing nature (e.g., antibiotics for pneumonia, lines 25 and 27; chemotherapy (implicit) for cancer, line 27).

Extending the frame used by FP to other medical conditions and their treatments, D is not only forcing FP to look at the weakness of her metaphorical argument; D is mainly shifting the communication to a different level, which does not seem to respect the emotions and desires previously exposed by FP. Indeed, because of D's argument, FP is (dialogically) pushed far away from her earlier arguments and worries (e.g., she never wanted to become a mom, lines 12, 14; she is afraid of having an unhealthy baby, line 20) and ends up agreeing with D's argument. To be more specific, the problem is not so much with the reasonableness of D's argument, which is not under discussion here (but see Rossi *et al.* 2017); what is problematic is the dialogical legitimacy in deciding to challenge a metaphorical frame that is used to describe personal values and worries, highly charged from an emotional and psychological point of view. That is, implicit persuasion and manipulation can introduce biases within the communication process thus putting the autonomy of patients at risk. In this sense, persuasion should be used legitimately, respecting the patients' viewpoints, and avoiding manipulation.

3. Persuasion and metaphors in shared decision-making

Persuasion is not sheer manipulation, even though it is not always easy to distinguish between the two. That is also the reason why the role played by persuasion and, above all, implicit persuasion is controversial within the field of health communication (Engelhardt *et al.* 2016, Powell & Partin 2013, Rubinelli 2013, Shaw & Elger 2013).

The use of persuasion strategies can be considered appropriate only if it safeguards patients' decisions and values. Adopting a patient-centered model of care, scholars have pointed out that mutual persuasion between patients and healthcare providers is what makes free communication different from manipulation, giving equal possibilities to patients and healthcare providers to argue their points of view (Barilan & Weintraub 2001, Labrie & Schulz 2014, Smith & Pettegrew 1986). Namely, a patient-centered model of care refers to the possibility of considering the involvement of patients in medical choices as a way to safeguard the patient's autonomy and freedom, as well as to ensure better clinical outcomes (Elwyn, Edwards, Kinnersley & Grol 2000, Roter & Hall 2006, Stewart *et al.* 2000). In this context, it has been argued that both providers and patients bring a different but equally legitimate perspective, with providers bringing medical knowledge and expertise and patients bringing their illness experience – which includes personal preferences and values concerning what medical treatments are most consonant with their life plan (Elwyn *et al.* 2012, Stewart *et al.* 2000, Street, Makoul, Arora & Epstein 2009). A decision-making process that takes into account these (sometimes different) viewpoints is what is known as shared decision-making in medicine. Elwyn and colleagues (2010: 971) use the following definition:

> Shared decision making is an approach where clinicians and patients make decisions together using the best available evidence. Patients are encouraged to think about the available screening, treatment, or management options and the likely benefits and harms of each so that they can communicate their preferences and help select the best course of action for them. Shared decision making respects patient autonomy and promotes patient engagement.

Considering this definition, it is even clearer why the ART case above does not seem to be ethically legitimate. Calling into question the metaphorical frame, the doctor was not providing further medically factual and relevant information that would have helped the patient to re-assess her convictions; that is, the doctor's argument was not free of value judgments, implying a delegitimization of the preferences and worries previously expressed by the patient.

3.1 Finding common ground through metaphors

Scholars have investigated how patients' perception of patient-centeredness is related to the shared decision-making process and affects the quality of care, understood both in terms of clinical outcome and improvement of patients' living conditions. A seminal study conducted by Stewart and colleagues (2000) showed the impact of two main factors in this process: first, an impact is obtained by how a problem is discussed and the patient's illness experience is explored; secondly, and most importantly, there is an impact of how discussion and agreement about treatment options are framed. Namely, finding common ground regarding management has a key role in allowing shared decision-making (and patient-centered medicine). As the authors note:

> Being patient centered does not mean that physicians abdicate control to the patient but rather that they find common ground in understanding the patients and more fully respond to their unique needs (Stewart *et al.* 2000: 797).

In every communicative interaction, finding common ground is a constraint on understanding and decision-making (Clark 1996, Stalnaker 2002). Metaphors are useful precisely where there is a gap in knowledge between speaker and hearer that needs to be filled: in such cases, interlocutors dynamically co-construct common knowledge emerging during the interaction; this dynamic process of co-construction is grounded in what interlocutors already share as their core common ground and accounts for the notion of emergent common ground.

A socio-cognitive approach to pragmatic inferences in intracultural and intercultural contexts proposes a dynamic and stratified view of common ground aiming at distinguishing between core common ground (the static, generalized, common repertoire of knowledge) and emergent common ground (the dynamic, actualized, and particularized contextual part of knowledge) (Kecskes 2008, Kecskes & Zhang 2009, 2013). Within this framework, metaphors can be used as tools promoting implicit arguments that are crucial in co-constructing the common ground and, therefore, also in reaching a shared understanding where an asymmetrical distribution of information is at stake. In asymmetrical relationships such as patient-provider interactions, where the distribution of knowledge and procedures is not shared by speakers, the use of an appropriate metaphorical frame could be useful to co-construct the (emergent) common ground (expressed through the target), hinged upon the already shared (core) common ground (expressed through the source) (Rossi 2016).

Two cases of metaphors detected in an Italian corpus of medical interviews between healthcare providers and patients with type 2 diabetes (Bigi 2014) are discussed to illustrate this point. The first example (Example 3) illustrates the use of a metaphorical expression by the patient.

> **Example 3**[6]
> Paziente: Poi ho notato che se mangio gli gnocchi, mi si svuota in fretta. A me piacciono tantissimo.
> Infermiera: Come si svuota in fretta?
> Paziente: Eh va giù, va giù.
>
> Patient: Then I have noticed that if I eat gnocchi, it empties itself quickly. I love gnocchi very much.
> Nurse: How does it empty itself quickly?
> Patient: Eh it goes down, goes down

By saying «si svuota in fretta» (*It empties itself quickly*) the patient is trying to explain what she thinks is clinically relevant evidence, *i.e.* when she eats gnocchi her glycaemia goes down, and this is not very good for her health quality. In this example, the metaphor seems to work as an attractor: the nurse fails to understand what the patient is saying and therefore asks for further clarification (*How does it empty itself quickly?*). However, the nurse understands that the patient is giving relevant information: the latter may be useful in finding a common ground between them and, consequently, in modifying the medical

[6] Example 3 was already discussed from a different perspective in Rossi & Macagno (2021) and Macagno & Rossi (2021).

recommendation and thus helping the patient to self-manage diabetes in a better way.

Example 4 concerns the use of a metaphorical frame introduced by a dietician and provides insights on how metaphors can be used by providers as tools to co-construct the common ground to reach a shared decision with the patients. More specifically, the dietician uses a metaphorical frame (detectable mainly in the expression "engaged couple") to explain the relationship between changes in weight and changes in diabetes compensation. By doing so, she is explaining what normally happens when people with diabetes gain weight, but she is also offering a persuasive argument about the acceptability of her medical recommendation (*i.e.*, to try losing weight).

> **Example 4**
> Dietologa: bisogna cercare di arrivare, avvicinarsi più che possiamo al peso ideale. Non aumentare. Perché generalmente diabete e peso viaggiano come due fidanzati, mano nella mano. Allora, se lei mi aumenta di peso, anche il diabete tende un pochino a salire.
>
> Dietician: you must try to reach, to get close to the ideal weight. Not gain weight. That's because diabetes and weight generally go hand in hand, like an engaged couple. So, if you gain weight, also diabetes tends to increase a bit. (*translation is mine*)

The dietician is trying to properly persuade the patient by framing a clinical problem through the use of a metaphor. The greater persuasive effect of metaphorical messages compared to the literal ones have been discussed in two recent meta-analyses (Sopory & Dillard 2002, Van Stee 2018). However, the problem of determining how patients understand these metaphors (Rossi & Macagno 2020), and which sources are more appropriate to use with which targets, remains. In this regard, Van Stee (2018) underlined that familiarity with the topic is important in generating persuasiveness, with a highly familiar target being generally more persuasive. Together with other studies (e.g., Hoeken, Swanepoel, Saal, & Jansen 2009), these results shed light on the importance of the background information/prior knowledge people must have at their disposal to properly exploit metaphors, gaining knowledge from them, and successfully assessing messages in which they are included.

3.2 Common ground and metaphors in intercultural contexts

The sharing of background information and prior knowledge is essential for understanding metaphors properly and exploiting them as an effective persuasion strategy in shared decision-making. The lack of a pre-existing common ground shared by the interlocutors may explain why the reconstruction of metaphorical meanings is sometimes made more difficult within intracultural interactions between patients and healthcare providers (Macagno & Rossi 2019).

The lack of common ground can be even more problematic in the contexts of intercultural and cross-cultural communication, where metaphors have been considered as potential sources of misunderstanding and communicative breakdowns (Kecskes 2006, Musolff 2014, Roberts, Moss, Wass, Sarangi, & Jones 2005, Sharifian 2014). Metaphor interpretation can become a problem when different cultures meet since the values employed within different languages and cultures can differ widely (Kövecses 2005, 2010, 2015, Musolff 2015). The impact of cultural differences on the interpretation of metaphors has been underscored in contexts of intercultural communication between native and non-native speakers, in which metaphor use was shown to cause comprehension difficulties (Littlemore, Chen, Koester, & Barnden 2011).

Intercultural communication seems to be deeply affected by the fact that interlocutors share scant common ground information related to the values characterizing the different linguistic and cultural communities to which they belong. This gap in cultural knowledge and common ground characterizing intercultural interactions might cause a mismatch in metaphorical understanding, setting the speaker and the hearer out along different interpretative paths. In these cases, the use of metaphors might be dangerous and ineffective (Rossi & Macagno 2021).

Despite these risks, metaphors might be positively exploited as a tool allowing the co-construction of the emergent common ground (Kecskes 2006), even in the case of intercultural and cross-cultural contexts. Metaphors, especially didactic and pictorial metaphors, may be exploited in intercultural and cross-cultural contexts to favour the finding of common ground between patients and healthcare providers.

In intracultural contexts, the use of pictorial metaphors in specialist discourses and medical didactics has been proved of great value (Bleakley 2017, Karska & Prażmo 2017). For example, it has been pointed out that metaphorical images are preferred over non-metaphorical images to illustrate medical concepts (Sánchez & Valenzuela 2019).

In the context of diabetes care, a conceptual metaphor exploiting the visual modality has been used within an intervention study to educate patients and foster their understanding of the clinical importance of three key metabolic markers (Hemoglobin A_1C, systolic Blood pressure, and low-density lipoprotein (LDL) Cholesterol values) (Naik, Teal, Rodriguez, & Haidet 2011) (fig. 1). This conceptual metaphor frames the daily activities related to diabetes self-management in terms of weather prediction; that is, it helps patients to interpret the prediction of the risk related to their ABC values by exploiting a weather metaphor. The authors of the study state:

> This metaphor, "predicting the weather", is especially effective in this setting because it uses low-literacy pictorial icons that have similar meaning across many populations, and because weather prediction is a widely understood concept for risk prediction (with the added benefit of conveying an understood sense of error and uncertainty in weather prediction). Participants in the current study appeared to comprehend and work very quickly with the information presented in Fig.1 and moved towards linking the ABCs to their daily activities (Naik Teal, Rodriguez, & Haidet 2011: 388).

Within this study, the weather metaphor proved to be effective with a range of different populations (White, African American, Hispanic or

Latino, and Other), with visual components probably also helping patients with low-literacy scores. Further studies are needed to extend these preliminary positive results to other metaphors and other populations, as well as to determine criteria helping to establish which metaphors are effective in intercultural and cross-cultural contexts.

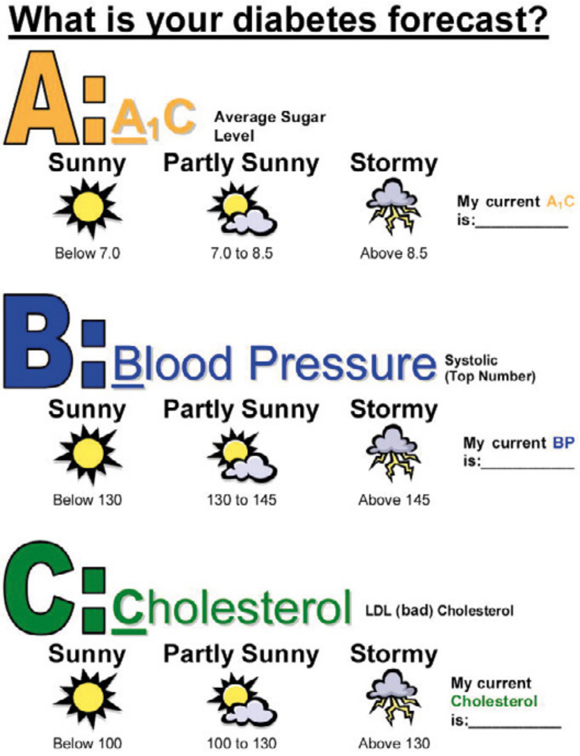


FIGURE 1
Conceptual metaphor used in the study by Naik Teal, Rodriguez, & Haidet 2011, p. 385.[7]

[7] The image can be retrieved at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116090/ (accessed on March 25th, 2023).

4. Conclusions

The paper has discussed metaphors as a tool to foster understanding and decision-making in the medical context. It explored this possibility by starting from the idea that metaphors are framing strategies containing implicit arguments that act as attractors of attention and have persuasive power. Metaphors are particularly useful where there is emotional and technical information that needs to be communicated, understood, and shared, which is the case of healthcare communication. However, their framing effect can also be dangerous: it might lead to ethical problems, introducing communicative biases, impacting the quality of care, and thus put patient safety at risk. It is therefore necessary to provide criteria to determine when/which metaphors are appropriate and why, extending current research also to intercultural and cross-cultural contexts. Finally, it showed how insights coming from the socio-cognitive approach developed within intercultural pragmatics can be used to shed light on some significant difficulties in the field of healthcare communication. Due to the growing presence of multilingual and intercultural contexts in our societies, future research should make more efforts to fruitfully integrate (intercultural) pragmatics in the context of health.

References

BARILAN Y. M. & WEINTRAUB M. (2001), «Persuasion as respect for persons: an alternative view of autonomy and of the limits of discourse», *The Journal of Medicine and Philosophy 26*(1), 13-33.

BIGI S. (2014), «Healthy Reasoning: The Role of Effective Argumentation for Enhancing Elderly Patients' Selfmanagement Abilities in Chronic Care», *in* G. Riva, P. Ajmone Marsan & C. Grassi (éds.), *Active Ageing and Healthy Living: A Human Centered Approach in Research and Innovation as Source of Quality of Life*, Amsterdam, IOS Press, 193-203.

BIGI S. (2018), «The role of argumentative practices within advice-seeking activity types. The case of the medical consultation», *Rivista Italiana Di Filosofia Del Linguaggio 12*(1), 42-52.

BLACK M. (1955), «Metaphor», *Proceedings of the Aristotelian Society, New Series*, *55*, 273-294.

BLACK M. (1979), «More about Metaphor. *Metaphor and Thought*, *31*, 19-43.

BLEAKLEY A. (2017), *Thinking with Metaphors in Medicine: The State of the Art*, New York, Routledge.

BORGHI L., LEONE D., POLI S., BECATTINI C., CHELO E., COSTA M., … VEGNI E. (2019), «Patient-centered communication, patient satisfaction, and retention in care in assisted reproductive technology visits», *Journal of Assisted Reproduction and Genetics 36*(6), 1135-1142.

BURGERS C., KONIJN E. A., & STEEN G. J. (2016), «Figurative Framing: Shaping Public Discourse Through Metaphor, Hyperbole, and Irony», *Communication Theory*, *26*(4), 410-430.

CAMERON L. (2003), *Metaphor in educational discourse*, London, Continuum.

CASARETT D., PICKARD A., FISHMAN J. M., ALEXANDER S. C., ARNOLD R. M., POLLAK K. I. & TULSKY J. A. (2010), «Can metaphors and analogies improve communication with seriously ill patients? », *Journal of Palliative Medicine 13*(3), 255-260.

CLARK H. (1996), *Using Language*, Cambridge, Cambridge University Press.

COSTELLO B. A. & ROBERTS F. (2001), «Medical Recommendations as Joint Social Practice», *Health Communication 13*(3), 241-260.

DEMJÉN Z. & SEMINO E. (2016), «Using metaphor in healthcare: Physical health», *in* E. Semino & Z. Demjén, *The Routledge Handbook of Metaphor and Language*, New York, Routledge, 385-399.

DOURY M. (2003), «L'évaluation des arguments dans les discours ordinaires: Le cas de l'accusation d'amalgame», *Langage et société* 105(3), 9-37.

ELWYN G, EDWARDS A., KINNERSLEY P. & GROL R. (2000), «Shared decision making and the concept of equipoise: the competences of involving patients in healthcare choices», *The British Journal of General Practice : The Journal of the Royal College of General Practitioners 50*(460), 892-899.

ELWYN G., FROSCH D., THOMSON R., JOSEPH-WILLIAMS N., LLOYD A., KINNERSLEY P., … BARRY M. (2012), «Shared decision making: a model for clinical practice», *Journal of General Internal Medicine 27*(10), 1361-1367.

ELWYN G., LAITNER S., COULTER A., WALKER E., WATSON P. & THOMSON R. (2010), «Implementing shared decision making in the NHS», *BMJ (Clinical Research Ed.)*, *341*, c5146.

ENGELHARDT E. G., PIETERSE A. H., VAN DER HOUT A., DE HAES H. J. C. J. M., KROEP J. R., QUARLES VAN UFFORD-MANNESSE P., … STIGGELBOUT A. M. (2016), «Use of implicit persuasion in decision making about adjuvant cancer treatment: A potential barrier to shared decision making», *European Journal of Cancer 66*, 55-66.

ENTMAN R. M. (1993), «Framing: Toward Clarification of a Fractured Paradigm», *Journal of Communication 43*(4), 51-58.

ERVAS F., GOLA E. & ROSSI M. G. (2018), «Argumentation as a Bridge Between Metaphor and Reasoning», *in* S. Oswald, T. Herman & J. Jacquin (éds.), *Argumentation and Language — Linguistic, Cognitive and Discursive Explorations*, New York, Springer International Publishing, 153-170.

FENG B., BELL R. A., JERANT A. F. & KRAVITZ R. L. (2011), «What do doctors say when prescribing medications?: An examination of medical recommendations from a communication perspective», *Health Communication 26*(3), 286-296.

GIBBS R. (1994), *The poetics of mind: figurative thought, language and understanding*, Cambridge, Cambridge University Press.

GIBBS R. & FRANKS H. (2002), «Embodied metaphor in women's narratives about their experiences with cancer», *Health Communication 14*(2), 139-165.

HARRINGTON K. J. (2012), «The use of metaphor in discourse about cancer: a review of the literature», *Clinical Journal of Oncology Nursing 16*(4), 408-412.

HAUSER D. J. & SCHWARZ N. (2014), «The War on Prevention: Bellicose Cancer Metaphors Hurt (Some) Prevention Intentions», *Personality and Social Psychology Bulletin 41*(1), 66-77.

HOEKEN H., SWANEPOEL P., SAAL E. & JANSEN C. (2009), «Using Message Form to Stimulate Conversations: The Case of Tropes», *Communication Theory 19*(1), 49-65.

JACOBI D. (1984), «Figures et figurabilité de la science dans des revues de vulgarisation», *Langages* (75), 23-42.

KARSKA K. & PRAŻMO E. (2017), «Didactic potential of metaphors used in medical discourse», *LingBaW*, *3*, 102-116.

KECSKES I. (2006), «Formulaic language in English Lingua Franca», *in* I. Kecskes & L. R. Horn (éds.), *Explorations in Pragmatics - Linguistic, Cognitive and Intercultural Aspects*, Berlin, De Gruyter, 1-28.

KECSKES I. (2008), «Dueling contexts: A dynamic model of meaning», *Journal of Pragmatics 40*(3), 385-406.

KECSKES, I. (2013), *Intercultural pragmatics*, Oxford, Oxford University Press.

KECSKES I. & ZHANG F. (2009), «Activating, seeking, and creating common ground: A socio-cognitive approach», *Pragmatics & Cognition 17*(2), 331-355.

KECSKES I. & ZHANG F. (2013), «On the Dynamic Relations Between Common Ground and Presupposition», *in* A. Capone, F. Lo Piparo & M. Carapezza (éds.), *Perspectives on Linguistic Pragmatics, Perspectives in Pragmatics, Philosophy & Psychology 2* Cham: Springer, 375-395.

KÖVECSES Z. (2005), *Metaphor in culture: Universality and variation*, Cambridge, Cambridge University Press.

KÖVECSES Z. (2010), «Metaphor, language, and culture», *DELTA: Documentação de Estudos Em Lingüística Teórica e Aplicada 26*(SPE), 739-757.

KÖVECSES Z. (2015), *Where Metaphors Come From: Reconsidering Context in Metaphor*, London, Oxford University Press.

KUNNEMAN M., GIONFRIDDO M. R., TOLOZA F. J. K., GÄRTNER F. R., SPENCER-BONILLA G., HARGRAVES I. G., … MONTORI V. M. (2019), «Humanistic communication in the evaluation of shared decision making: A systematic review», *Patient Education and Counseling 102*(3), 452-466.

LABRIE N. & SCHULZ P. J. (2014), «Does argumentation matter? A systematic literature review on the role of argumentation in doctor-patient communication», *Health Communication 29*(10), 996- 1008.

LAKOFF G. & JOHNSON M. (1980), *Metaphors we live by*, Chicago, University of Chicago Press.

LEONE D., BORGHI L., DEL NEGRO S., BECATTINI C., CHELO E., COSTA M., … VEGNI E. (2018), «Doctor–couple communication during assisted reproductive technology visits», *Human Reproduction 33*(5), 877-886.

LITTLEMORE J., CHEN P. T., KOESTER A. & BARNDEN J. (2011), «Difficulties in metaphor comprehension faced by international students whose first language is not English», *Applied Linguistics 32*(4), 408-429.

MACAGNO F. & ROSSI M. G. (2019), «Metaphors and problematic understanding in chronic care communication», *Journal of Pragmatics 151*, 103-117.

MACAGNO F. & ROSSI M. G. (2021), «The communicative functions of metaphors between explanation and persuasion», *in* F. Macagno & A. Capone (éds.), *Inquiries in philosophical pragmatics. Theoretical developments*, Cham, Switzerland, Springer, 171-191.

MACAGNO F. & ZAVATTA B. (2014), «Reconstructing Metaphorical Meaning. *Argumentation*», *28*(4), 453-488.

MUSOLFF A. (2014), «Metaphors: Sources for intercultural misunderstanding?», *International Journal of Language and Culture 1*(11), 42-59.

MUSOLFF A. (2015), «Metaphor interpretation and cultural linguistics», *Language and Semiotic Studies 1*(3), 35-51.

NAIK A. D., TEAL C. R., RODRIGUEZ E. & HAIDET P. (2011), «Knowing the ABCs: a comparative effectiveness study of two methods of diabetes education», *Patient Education and Counseling 85*(3), 383-389.

ORTONY A. (1975). «Why Metaphors Are Necessary and Not Just Nice», *Educational Theory 25*(1), 45-53.

PERELMAN C. & OLBRECHTS-TYTECA L. (1969), *The New Rhetoric: A treatise on argumentation*, Notre Dame, Ind., University of Notre Dame Press.

PETTY R. & CACIOPPO J. (1986), «The Elaboration Likelihood Model of Persuasion», *Advances in Experimental Social Psychology 19*, 123-205.

POWELL A. A. & PARTIN M. R. (2013), «The Role of Persuasion», *JAMA*, *310*(6), 646-647.

ROBERTS C., MOSS B., WASS V., SARANGI S. & JONES R. (2005), «Misunderstandings: A qualitative study of primary care consultations in multilingual settings, and educational implications», *Medical Education 39*(5), 465-475.

ROSSI M.G. (2016), «Metaphors for patient education. A pragmatic-argumentative approach applying to the case of diabetes care», *Rivista Italiana Di Filosofia Del Linguaggio 10*(2), 34-48.

ROSSI M.G., LEONE D., & BIGI S. (2017), «The ethical convenience of non-neutrality in medical encounters: Argumentative instruments for healthcare providers», *Teoria 37*(2), 139-157.

ROSSI M. G. & MACAGNO F. (2020), «Coding Problematic Understanding in Patient–provider Interactions», *Health Communication 35*(12), 1487-1496.

ROSSI M. G. & MACAGNO F. (2021), «Intercultural pragmatics in healthcare communication: An overview of the field», *in* I. Kecskes (éd.), *Cambridge Handbook of Intercultural Pragmatics*, Cambridge, Cambridge University Press.

ROSSI M. G. & MENICHETTI J. (2019). «Health and participation: Facebook as an educational tool for engaging patients», *Sistemi Intelligenti* 31(3), 570-599.

ROTER D. & HALL J. A. (2006), *Doctors talking with patients/patients talking with doctors: improving communication in medical visits*, Westport, Greenwood Publishing Group.

RUBINELLI S. (2013), «Rational versus unreasonable persuasion in doctor-patient communication: a normative account», *Patient Education and Counseling* 92(3), 296-301.

SÁNCHEZ M. T. & VALENZUELA A. C. (2019), «Visual metaphors in medical knowledge representation», *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *17*(0). [consulté le 25-01-2021, https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/464]

SCHERER A. M., SCHERER L. D. & FAGERLIN A. (2014), «Getting Ahead of Illness: Using Metaphors to Influence Medical Decision Making», *Medical Decision Making 35*(1), 37-45.

SEMINO E. (2008), *Metaphor in discourse* , Cambridge, Cambridge University Press.

SEMINO E., DEMJÉN Z., DEMMEN J., KOLLER V., PAYNE S., HARDIE A. & RAYSON P. (2015), «The online use of Violence and Journey metaphors by patients with cancer, as compared with health professionals: a mixed methods study», *BMJ Supportive & Palliative Care*. [consulté le 25-01-2021, https://spcare.bmj.com/content/7/1/60 ; DOI : 10.1136/bmjspcare-2014-000785]

SHARIFIAN F. (2014), «Conceptual metaphor in intercultural communication between speakers of Aboriginal English and Australian English», *in* A. Musolff, F. MacArthur, & G. Pagani (éds.), *Metaphor and intercultural communication*, London, Bloomsbury, 117-129.

SHAW D. & ELGER B. (2013), «Evidence-Based Persuasion: An Ethical ImperativeThe Ethics of Evidence-Based Persuasion», *JAMA 309*(16), 1689-1690.

SMITH D. H. & PETTEGREW L. S. (1986), «Mutual persuasion as a model for doctor-patient communication», *Theoretical Medicine* 7(2), 127-146.

SOPORY P. (2017), «Metaphor in Health and Risk Communication», *in* R. L. Parrott (ed.), *The Oxford Encyclopedia of Health and Risk Message Design and Processing*, London, Oxford University Press, 188-213.

SOPORY P. & DILLARD J. P. (2002), «The Persuasive Effects of Metaphor a Meta-Analysis», *Human Communication Research 28*(3), 382-419.

STAHL B. (1998), «Le consentement à l'acte médical», *Revue juridique de l'Ouest 11*(3), 289-314.

STALNAKER R. (2002), «Common Ground», *Linguistics and Philosophy 25*, 701-721.

STEWART M., BROWN J. B., DONNER A., MCWHINNEY I. R., OATES J., WESTON W. W., & JORDAN J. (2000), «The impact of patient-centered care on outcomes», *The Journal of Family Practice 49*(9), 796-804.

STREET R. L. J., MAKOUL G., ARORA N. K. & EPSTEIN R. M. (2009), «How does communication heal? Pathways linking clinician-patient communication to health outcomes», *Patient Education and Counseling 74*(3), 295-301.

VAN STEE S. K. (2018), «Meta-Analysis of the Persuasive Effects of Metaphorical vs. Literal Messages», *Communication Studies 69*(5), 545-566.

Simplicity of What?
A Case Study from Generative Linguistics

Giulia Terzian and María Inés Corbalán

Terzian, G. & Corbalán, M. I. (2021). Simplicity of what?
A case study from generative linguistics.
Synthese, **198, 9427-9452.**
https://doi.org/10.1007/s11229-020-02649-5

1. Introduction

Simplicity[1] is widely hailed across science and philosophy as a desirable trait of our theories, models, explanations, etc. Generative linguistics is no exception, on the contrary going so far as to elevate simplicity to the status of high priority research goal.[2] It is therefore striking, given the purported centrality of this notion, that generativists have yet to offer satisfactory answers to the fundamental questions of how simplicity is to be defined, measured, traded-off and–above all–justified. As we will argue, the latter worry in particular becomes even more pressing under the recent Minimalist Program (MP), which is predicated on the idea that simplicity is a fundamental and defining feature of the human language faculty, a key ingredient in linguistic explanation, and a prominent theoretical constraint (Chomsky 1995). In order to back up these claims, we begin by reviewing what we see as the most salient junctures in generative conceptualizations of simplicity, in Sect. 2.[3] Among other things, this exercise will reveal that there continues to be a good deal of ambiguity concerning the alleged bearer(s) of this notion, thus explaining

[1] And cognate notions such as parsimony, economy, elegance, naturalness, beauty, etc. Unless explicitly stated, any mention of simplicity should be interpreted as shorthand for 'simplicity and cognate notions'. To be absolutely clear, we are not claiming that any or all of these notions are equivalent and interchangeable; we are merely using 'simplicity' as an umbrella term for the sake of discursive fluidity. Readers will find this important clarification borne out in the forthcoming discussion.

[2] Generative linguistics is typically construed as a branch of cognitive science, insofar as its chief concern is the study of the human language faculty, and the latter is a component of our cognitive system. The point of this extremely rough characterisation is to distinguish the study of language as an internal state of a biological organ from language as a socio-cultural object.

[3] We wish to emphasise that this will not merely consist of a regurgitated version of agreed-upon facts: the chaotic, disconnected nature of the literature makes this near impossible. Sect. 2 is the result of our own laborious reconstruction of the history of simplicity in generative linguistics. We are not claiming that this is the best or only such reconstruction (for a critical analysis of generativism from the perspective of linguistic historiography, see Kertész 2010); nor that the stages identified therein are entirely clearcut or conceptually isolated from one another. Indeed, our discussion explicitly marks the continuities underwriting the evolving generative conceptions of simplicity.

the widespread and well-documented (and otherwise unwarranted) expectation that the ontological and theoretical notions of simplicity should converge (Sect. 3). The second and main part of the paper is devoted to showing that the issues of justification and convergence become much more tractable as long as generativists embrace a more naturalistic methodology; importantly, our proposal will be conciliatory rather than antagonistic.[4] We make this case in Sects. 4–6, by examining the notion of simplicity through the lens of a pair of recent debates in cognitive science and philosophy–respectively, on domain-general cognitive biases and on scientific understanding. Among other things we show that in its object-level capacity, simplicity is much more plausibly construed as a derived (or inherited) vs. intrinsic property of the language faculty. Moreover, we argue that minimalist appeals to simplicity as a theoretical value can be justified–as long as minimalists themselves adopt a more flexible perspective of the aims of scientific inquiry on the one hand, and of which epistemic vehicles can further such aims on the other. Section 7 concludes.

2. Simplicity in generative linguistics: a bird's-eye view

2.1. Simplicity double-act: theory selection and grammar selection

Up until the mid-50s, the main goal of generative linguistics was to arrive at a descriptively adequate characterisation of human languages (Chomsky 1955, 1957). Simplifying greatly, this amounted to a two-fold task: formulating grammars– understood here as systems of rules–underlying existing languages, and producing a general *theory*

[4] We use the term 'naturalism' to refer to the methodological approach, or attitude, that explicitly encourages and values a frank dialogue between philosophy and the sciences, and furthermore acknowledges that such dialogue may require philosophers to defer to scientists' expertise; see e.g. Nersessian (1987), Ankeny et al. (2011) and Soler et al. (2014). This is a salient clarification given that the term 'naturalism' is sometimes appropriated by generativists to refer to the so-called 'Galilean stance' (cf. Sect. 3).

of grammar. Accordingly, up until this point simplicity appeared in a purely methodological capacity, shaping the search for 'best' theory into the search for the theory of grammar that is 'simplest': more unified, containing fewer and shorter rules, and fewer symbols.[5]

The first salient juncture coincides with the explanatory turn of the 60s, as generativists direct their attention to the question of *how* individual linguistic agents learn, or acquire, (their native) language.[6] Loosely put, the idea in these early stages of linguistic theory is that at birth, a speaker's native language is underdetermined by the available evidence (external linguistic stimuli); language acquisition comes about as the speaker (or rather the speaker's linguistic module) 'chooses' among possible grammars, eventually settling on the correct one. But how does our cognitive apparatus complete such a task, given the infinite size of this class? To obviate this difficulty,

> For the construction of a reasonable acquisition model, it is necessary to reduce the class of attainable grammars compatible with given primary linguistic data to the point where selection among them can be made by a formal evaluation measure. (Chomsky 1965, p. 35)

Crucially, the task of 'reducing the class of attainable grammars' is now explicitly ascribed to the human language faculty. More specifically, on this early explanatory account it is postulated that humans are genetically endowed with a rich 'universal grammar'–consisting of more or less abstract rules–which therefore curtails the space of 'attainable grammars compatible with given primary linguistic data.' This posited universal grammar thus turns language acquisition from an impossible to a feasible task; at the same time, it is not thought to achieve a definitive reduction of the space of possible grammars. That is, it is

[5] The origins of this grammar-specific simplicity criterion are found in (Chomsky 1951, p. 6): "the criteria of simplicity are as follows: that the shorter grammar is the simpler, and that among equally short grammars, the simplest is that in which the average length of derivation of sentences is least."

[6] At this stage, the terms 'learn' and 'acquire' were used fairly loosely and interchangeably.

still thought that there can be more than one descriptively adequate grammar for a given language; and that given two descriptively adequate grammars, (part of) the role of the language faculty is to provide the procedure for selecting the 'correct' one. Chomsky then postulates that simplicity enters this very selection procedure; put differently, and only slightly more precisely, it is thought that some sort of simplicity metric (or ranking, or evaluation) is part of the actual process of language acquisition:

> Here in outline is the device Chomsky used in the mid-1960s to make sense of how the child's mind automatically 'selects' grammar X as opposed to Y–that is, learns X as opposed to Y, given data D. Think of X and Y as sets of rules, both candidates as descriptions of language L or, more carefully, of the data available to the child's mind. Which [...] should the child's mind choose? Introduce now an 'internal' simplicity measure: rule set X is better than Y to the extent that X has fewer rules than Y. (Chomsky 2009, p. 28).

Thus, simplicity makes its first 'double' appearance, in an object-level as well as a theoretical capacity. Moreover, the internal notion is thought to play a prominent role in language acquisition, roughly in analogy to the way that supra-empirical criteria intervene in underdetermination scenarios.[7]

2.2. Simplicity internalised: from internal metric to innate endowment

The next key turn comes about as the explanatory question is gradually sharpened into the formulation now known as Plato's Problem: How do children acquire language given the poverty of data initially available to them?

[7] See also Sober (1975, Ch. 2) and Sober (1978).

A little more specifically, foremost on the research agenda at this stage is the challenge of explaining the following observed facts about linguistic acquisition and competence:[8]

> (P1) the homogeneity of language acquisition within and across linguistic communities;
> (P2) the relatively short time it takes children to acquire their native language, given the poverty of input data;
> (P3) the vast diversity of languages.

Ultimately, generative efforts to account for (P1)–(P3) crystallised into the so-called Principles & Parameters framework (Chomsky 1981). The P&P model paints the following picture of the human language faculty (FL).[9] In its initial state (i.e., when we are born) FL is genetically equipped with two types of resources: a set of universal principles and a set of 2-valued parametrized principles. In this initial state–known as Universal Grammar (UG)–the parametrized principles are 'switched off'; the classic analogy invoked in the literature is of a dormant switchboard.[10] Prompted by linguistic stimuli from the environment, FL 'sets' the value of these parameters. Language acquisition is what happens as more and more parameters are set, as a result of an optimal interaction between FL and the linguistic environment. Once all parameters have been fixed, the (idealized) native speaker has achieved linguistic competence, i.e. language acquisition is complete. Crucially, (P1)–(P3) receive an elegant and seemingly plausible explanation by the lights of this model.

[8] Linguistic competence is sometimes described as a kind of knowledge (of the grammar of the speaker's native language), although it remains an open question just what sort of knowledge might be at stake. For instance, Chomsky categorically and convincingly rules out that it be identified with propositional knowledge (knowledge-that), and suggests it is a kind of tacit knowledge.

[9] The literature refers to P&P interchangeably as a model, a theory, an approach, a framework, or even a program.

[10] UG is sometimes referred to as "the theory of the initial state of the language faculty" (Chomsky 1995, p. 12), or "the theory of the biological endowment of [FL]" (Chomsky 1995, p. vii); and sometimes simply as "The biological equipment that makes language acquisition possible" (Boeckx 2010, p. 486). In this context the term 'theory' seems to be used rather loosely, then, to denote a cluster of constraints that pick out UG. See also Sect. 3.

Notice however the absence of any explicit reference to simplicity in the foregoing, either as a theoretical property or as an internal feature of FL; yet there is no doubt that generativists continue to entertain both assumptions. A plausible explanation is that P&P is thought to embody both constraints, thus foregoing the need to make either explicit. How so? We suggest the following interpretations. First, P&P is a simpler theoretical construction compared to its predecessor, in three respects:

- Ontological parsimony: a small number of abstract principles and 2-valued parameters replace a complex structure of specific rules;
- Unification: UG is universal in a stronger sense than its lower-case predecessor;
- Explanatory power: language acquisition is now a (comparatively) low-complexity task.

Secondly, FL itself instantiates three kinds of simplicity on the P&P account:

- Elegance and unification: the constituents of UG are fewer and highly abstract;
- Economy: FL operates more efficiently and with fewer resources.

2.3. From Plato's problem to Darwin's problem

The final turn coincides with the birth of the recent Minimalist Program (Chomsky 1993, 1995). MP takes as premises that the generative enterprise, up to and including P&P, has successfully addressed both the descriptive challenge (by identifying the particular grammars underpinning individual languages) and a first layer of the explanatory challenge (by producing a model of human language acquisition). As we've just seen, the gist of the latter is that FL develops or 'grows' from an initial, universal state–UG–to a steady state–the individual language/ grammar –, prompted by environmental linguistic stimuli.

Minimalism explicitly seeks to address a second layer of the explanatory challenge, sometimes described as the challenge of arriving at a *principled* explanation of the properties of FL.[11] More specifically, informing the minimalist research agenda are the following questions:

(M1) Exactly how does FL work?
(M2) Why does it have the properties it has?
(M3) How could FL have evolved?

MP's key conjecture is that FL is a cognitive module that interacts with nearby modules (the sensori-motor and the conceptual-intensional systems) in an optimal way. This is the so-called Strong Minimalist Thesis:

(SMT) FL is an optimal solution to the interface conditions imposed by the conceptual-intensional and sensori-motor systems.[12]

Importantly, minimalist attempts to substantiate SMT rely heavily (and once again explicitly) on two notions of simplicity, one external and one internal.[13] In fact, contrary to the official party line we find that extant discussions underwrite a more fine-grained taxonomy of simplicity:

- An external notion, labeled methodological economy (MS). This is the familiar– imprecisely defined–theoretical value, guiding linguistic inquiry (*qua* scientific inquiry).
- Two internal notions, typically lumped together under the label of ontological or substantive economy.

[11] Several authors have noted that early hints of this idea can be traced all the way back to early generative writings; that is, the idea that generative linguistics should aspire to one day achieve this sort of explanatory depth far predates MP itself. For further discussion of the conceptual continuities spanning generative history, see e.g. Freidin and Lasnik (2011), as well as Boeckx (2006).

[12] A more recent conjecture is that FL is an optimal solution only, or primarily, to C-I design specifications; see e.g. Chomsky (2007).

[13] The two types or levels of simplicity underlying MP are also discussed by Boeckx (2006) in terms of elegance and beauty, respectively. Thus, he writes that "Work within minimalism seeks to develop beautiful theories" (2006, p. 120) where a 'beautiful theory', as we understand it, is one that offers a satisfactory principled explanation of linguistic phenomena, as required by (M2) above.

- Procedural simplicity (PS): FL operations are subject to a number of economy constraints on derivations and on representations.[14]
- Ontological simplicity (OS): UG is ontologically parsimonious, sparse, non-redundant.

This is MP's main gamble: that *FL is both procedurally and ontologically simple*. By way of investigating this conjecture, minimalist inquiry has largely focused on re-examining extant linguistic accounts by the lights of MS, PS and OS. To the extent that "Minimalist considerations motivate rethinking and replacing [previously accepted] assumptions and technical machinery" (Hornstein et al. 2005, p. xii) this can be seen as an attempt to address (M1). More recently, minimalists have turned their attention to (M2)–the demand for a principled explanation of FL-properties–and (M3)–known as Darwin's Problem (Boeckx 2009). We'll briefly expand on these in turn.

Once again simplifying greatly, we may see attempts to address (M2) as guided by the 'third-factor hypothesis': that at least some and perhaps most properties of FL may derive from, and be explained by "even more general, perhaps "language-external" principles" (Chomsky 2004, p. 24). This idea stems from Chomsky's suggestion that

> the growth of language in the individual is determined by the interaction of three factors: (a) genetic endowment; (b) experience; and (c) general principles not specific to the language faculty. (Al-Mutairi 2014, p. 73)

What might these principles be? Beyond the fact that they are non-domain-specific, universal, and language-external, opinions on this

[14] Economy conditions on derivations and representations guarantee that the latter are optimal or (computationally) efficient in some—more or less precisely specified—sense. Various such conditions have been proposed in the minimalist literature, including e.g.: Inclusiveness; Shortest Move; Last Resort; Procrastinate; Greed; Enlightened Self-interest; No-tampering; Full Interpretation. Details don't really matter here; the general idea is that grammars are organized frugally to maximise resources. For a (very) critical discussion of economy principles proposed under MP, see Lappin et al. (2000).

matter diverge.[15] We are more interested in the fact that the hypothesis marks a fundamental shift in the allocation of explanatory burden. Recall the cardinal hypothesis of P&P: that UG–our genetically determined linguistic endowment–is rich enough to bear the explanatory bulk of language acquisition (and the workings of FL, more generally). By contrast, under MP it is thought that

> a "principled explanation" of the language faculty and its properties may be achieved by "shifting the burden of explanation from the first factor [...] to the third factor" (Chomsky 2005, p. 9). (Al-Mutairi 2014, p. 75)

Crucially, the rationale for such a shift comes from the generative community's more recent concern over reconciling models of FL with evolutionary theory. For, while P&P offers an attractive answer to Plato's Problem as a result of countenancing a rich, genetically encoded UG, this very assumption makes it problematic from an evolutionary perspective–particularly given the relatively short time that language has 'been around' (less than 100,000 years by most estimates).[16] This is

[15] The term 'third factor' is introduced by Chomsky (2005) (although the germ of the idea may have already been present in 1965), and characterised as consisting of "Principles not specific to the faculty of language," including "(a) principles of data analysis that might be used in language acquisition and other domains; (b) principles of structural architecture and developmental constraints [...] including principles of efficient computation" (Chomsky 2005, p. 6). Elsewhere, Chomsky occasionally seems to think that principles of type (b) might include general laws of physics. We'll disregard the latter interpretation: partly because it is comparatively underdeveloped, and partly because the former is more attuned with the focus of our discussion.

[16] In this connection, Boeckx draws attention to certain similarities between the acquisition and evolution challenges. Both raise a problem of reconciling a complex phenomenon—respectively, language acquisition at the individual level and evolution of FL at the species level—with the strict temporal constraints to which the phenomenon is subject. This observation in turn motivates adopting analogous solution strategies, Boeckx suggests: "the way we should try to address and solve [the evolution challenge] is to do exactly what we have done for Plato's Problem, namely to [...] make sure that the thing that has to evolve is actually fairly simple" (2009, p. 47). We merely add that whether the proposed approach—namely, to aim for parallel solutions based on certain similarities between the respective problems—will work is ultimately an (as yet open) empirical matter, and certainly not one we are equipped to pronounce ourselves on.

Darwin's Problem. In response, minimalists have adopted a two-pronged simplicity-based strategy, devised to ease the evolutionary pressure on FL and thus avoid having to posit 'multiple miracles': on the one hand, shift the burden of explanation from the first to the third factor; on the other, seek to 'empty' UG as much as possible, either by eliminating entities outright or by reducing them to a thinner and more fundamental ontological basis.

The picture that emerges from the foregoing (lamentably brief) overview could be described at once as dynamical and volatile. We've seen the notion of simplicity occupy a central role throughout the history of generative inquiry, albeit under rather changeable guises. In particular, we saw it double up as theory-level and object-level property fairly early on, before mutating further still–most recently, into what we have labelled PS and OS–in this latter capacity. What we have not seen–what is remarkably absent from the literature and not just our overview–is a corresponding, parallel narrative as to why we should take these simplicity ascriptions at face value. This is true not just of theory-level simplicity claims, for which robust justifications are notoriously hard to pin down in general. It is also and much more pointedly true of their object-level counterparts. As noted at the outset of the paper, this is a puzzling situation given the centrality of the idea that simplicity is a property of FL, both throughout generative history and most explicitly under MP. Indeed, in light of this latter fact the lack of a solid justificatory basis for either kind of simplicity claim becomes a legitimate and serious concern. Happily, we think there is a way to mitigate both worries, as we'll see in Sects. 4–6. Before we do so, the next section briefly expands on an additional important confounding factor in generative discussions of simplicity, witnessing a sustained conflation between theory- and object-level notions on the one hand, and an expectation that the two should converge on the other.

3. Galileo meets Ockham: the purported convergence of simplicities

Patently, simplicity concerns have been a constant fixture in the development of generative linguistics. By contrast, the interpretation of this notion has fluctuated considerably from one framework to the next, and sometimes within one and the same framework. More worryingly, discussions of simplicity are mired in at least one important sort of ambiguity, between ascriptions of simplicity to the object of study and to linguistic theory itself. A representative example of this sort of confusion is found in the following passage:

> To repeat, minimalism is a project: to see just how well designed the faculty of language is, given what we know about it. It's quite conceivable that it has design flaws, a conclusion we might come to by realizing that the best accounts contain a certain unavoidable redundancy or inelegance. (Hornstein et al. 2005, p. 14)

In fact, the conflation of theory- and object-simplicity is but one instance of a more general trend, within the generative community, of failing to disambiguate between theory and object *simpliciter*. Particularly notable instances of this tendency are the notions of 'grammar' (cf. 2.2) and, later on, UG. Thus, for instance, UG is described simultaneously as an object of linguistic inquiry–specifically, the system of universal constraints that constitute our innate linguistic endowment–and as the theory of that same object–i.e. the theory of the initial state of FL. This poses a non-trivial interpretation problem, for instance when it comes to understanding the linguist's directive to 'rethink the structure of UG', or 'minimise UG'.[17]

Acknowledging this conflationary habit affords us an intuitive grip on the minimalist expectation that theory- and object-simplicity

[17] Similarly, in the discipline's early days the term '(generative) grammar' was used ambiguously to refer both to the linguist's object of study (i.e. a grammar for a particular language), and the linguist's theory (i.e. the theory of generative grammar).

should converge. We suggest that this convergence assumption can be further unpacked in terms of the following explanatory factors: (E1) a largely implicit commitment to a strong form of (semantic scientific) realism, (E2) a commitment to a metaphysical thesis according to which the world is simple (known in generative circles as the Galilean stance, or style), (E3) a commitment to a 'naturalist' stance according to which 'language should be studied in the same way as any other aspects of the natural world' (Al-Mutairi 2014, p. 34), (E4) a commitment to the 'Occamist urge to explain with only the lowest number of assumptions' (Boeckx 2010, p. 494), (E5) a failure to clearly distinguish between (E1)–(E4).

Illustrations of (E1)–(E5) are anything but difficult to find in the literature. Here are just a few representative passages:

> We construct explanatory theories as best we can, taking as real whatever is postulated in the best theories we can devise (because there is no other relevant notion of 'real'), seeking unification with studies of other aspects of the world. (Chomsky 1996, p. 35) (as cited in (Smith and Allott 2016, p. 204))

> [What] further properties of language would SMT suggest? One is a case of Occam's razor: linguistic levels should not be multiplied beyond necessity, taking this now to be a principle of nature, not methodology, much as Galileo insisted and a driving theme in the natural sciences ever since. (Chomsky 2007, p. 16)

> [The] Galilean style [...] is the central aspect of the methodology of generative grammar. [...] The Galilean program is thus guided by the ontological principle that "nature is perfect and simple, and creates nothing in vain" [...]. This outlook is exactly the one taken by minimalist linguists. [...] The road to Galilean science is to study the simplest system possible [...]. (Boeckx 2010, p. 498)

> Without adhering to the Galilean style, without the strongest possible emphasis on simplicity in language (the strongest minimalist thesis), it is hard to imagine how we might ever make sense of the properties of FL. (Boeckx 2010, p. 501)

Notice the no-miracle flavour of the last quote; paraphrased from context, it amounts to the following: If FL weren't as MP describes it, (i) the success of MP would be a miracle and (ii) the evolution of FL would require multiple miracles. Interestingly, this parallels the argumentative strategy employed in justifications of a rich, innate UG (cf. also footnote 16). Paraphrasing from Al-Mutairi (2014) (and his paraphrase of Chomsky): Factor I must be non-empty ('*something* must be special to language') or else language acquisition would be a miracle; Factor III must be non-empty or else language evolution would be a miracle.

The foregoing sections have sought to unearth the many faces of simplicity in generative linguistics. Perhaps the most salient aspect of the resulting picture is a persistent and *indiscriminate* pull towards simplicity–an entrenched belief that simplicity colours both theory and object of study–that sits on a shaky foundation, captured by (E1)–(E5) above. In light of these facts, it is therefore hardly surprising that justification questions have been largely overlooked. In the next sections, we offer the minimalist a way out.[18]

[18] One of the referees remarked that, 25 years down the line, the minimalist community's interest in the original research program (i.e. arriving at an understanding of FL that goes beyond 'mere' explanatory adequacy, which would require inter alia seriously investigating the third factor hypothesis) has gradually dwindled, as minimalists have returned to more parochial (descriptive and explanatory) work on particular languages (or: grammars). The referee thus wondered whether there is still a minimalist community 'out there' that could fill the role of our audience. Put differently, who is 'the minimalist' we are purportedly addressing? Our response to this interesting observation is threefold. On the one hand, if the community has indeed drifted away from the program's original research goals, then—seeing as several of these goals have yet to be met—that is all the more reason to put these suggestions to them. We don't for a moment presume that our paper could be so impactful as to single-handedly turn the minimalist tide, of course. Still—and here is our second point—there are at the very least scattered individual researchers out there who remain invested in the original questions. However few in number, they are a worthy audience for this and like-minded papers. Finally, as we write in Sect. 6 we are in fact addressing more than just 'the minimalist'. To borrow

4. Taking the third-factor hypothesis to the next level

We see the rise in prominence of the third-factor hypothesis as one of the most promising aspects of recent minimalist inquiry. At the same time, it is our impression that its significance and potential ramifications have thus far been under-appreciated by the minimalist community.[19] In large part, this is because generative linguistics has not quite lived up to its own self-identification as a branch of cognitive science, at least insofar as it has foregone substantive engagement with said discipline. In this section we make a case for the importance of a collaborative dialogue between linguistics and cognitive science: not just for the sake of honouring the former's naturalistic commitment (although this would be a good enough reason by itself); but also, more pointedly, as a way to address and mitigate the justification worry with respect to object-simplicity claims. In light of a cluster of well-supported findings in cognitive science, we'll see, the long-standing generativist 'hunch' that FL is in some sense simple stands a good chance of being vindicated.

To see how, recall first that minimalists have sought to substantiate SMT by placing a premium on OS as a guide to constructing models of FL (Sect. 2.3). Such models thus witness a reduction of both the innate and the domain-specific content previously assumed to be part of UG. Moreover, while the implementation of OS sometimes results in the outright elimination of entities from UG, more often it leads to a relocation of content, either from UG to other cognitive systems (third factor), or from UG to the environment (second factor), or both. Crucially,

1. at least part of the content relocated to other cognitive systems consists of PS constraints;

Kitcher's (2019) terminology, we hope to address the Scientists (cognitive scientists as well as generativists), the Philosophers (especially those interested in the role of aesthetic values in scientific contexts), and last but not least the Interested Citizens (or more prosaically, any and all those who recognise the value of the sort of interdisciplinary investigation we are promoting).

[19] With (isolated) exceptions, of course. As we'll see below, some minimalists (e.g. Boeckx 2016) have embraced the methodological shift prompted by the third-factor hypothesis more explicitly than others.

> 2. to the extent that SMT is true, content that is relocated to other cognitive systems is still 'part of'–or accessible to–FL.[20]

More plainly: taken together, SMT and the third-factor hypothesis entail (among other things) that simplicity is no longer a domain-specific property of UG, but rather *a domain-general cognitive feature*. Oddly, minimalists have largely downplayed or even ignored the ramifications of this fact, nor have they ventured to seek its corroboration (or correction) from empirical evidence.[21]

[20] The widely cited paper by Hauser et al. introduces a distinction between the language faculty in a broad sense (FLB)—which includes the interface systems—and in a narrow sense (FLN)—comprising just the recursive computational system. The authors' hypothesis, which enjoyed a rapid uptake within the generativist community, is that "FLN [...] is the only uniquely human component of the faculty of language" (2002, p. 1569), whereas FLB is (plausibly) shared with other species. In other words, on this hypothesis: some language-specific content exists; and all (and only) such content is confined to FLN.

[21] Once again, we do not claim that *no* minimalist has taken the third factor hypothesis and its consequences seriously; only that this is true of the community, at a programmatic level. For instance, one of the reviewers pointed out that Boeckx (2014b, 2016) clearly favours the idea that third factor content—under the heading of FLB—could play a non-trivial role in language acquisition, and indeed engages with literature from cognitive science to discuss specific (possible) illustrations of such content. Indeed, in more than one place Boeckx is explicit about his commitment to the 'biolinguistic enterprise', understood as "The road leading theoretical linguistics beyond explanatory adequacy, that is, towards a naturalistic, biologically grounded, better-integrated cognitive science of the language faculty" (2014a, p. 1). In this respect, we are very much on the same page as Boeckx: we too are pushing for genuinely interdisciplinary collaboration between the linguistic and cognitive science communities; we too regard the third factor hypothesis as very much worth investigating from an integrated cognitive perspective. What sets apart our proposal from his hinges on our respective foci. Put briefly and certainly not exhaustively, in his more recent work Boeckx seeks to explore the consequences of the third factor hypothesis on what we might term a more general scale: as a means to (re-) address Plato's Problem by effecting a 'fairer' redistribution of the explanatory burden between FLN (comprising perhaps no more than a universal, invariant computational system) and FLB (comprising now also domain-general learning mechanisms). Our focus in this subsection is in a sense much narrower, and informed by somewhat different premises (something not altogether surprising since, as philosophers, we are 'outsiders' to the linguistic debate). We are interested specifically in the 'third factor consequences' for the (internal) notion of simplicity; in a sense, we want to see 'what happens' to this central notion if and once an interdisciplinary

We think such evidence can be found in recent empirical studies conducted by cognitive scientists of different ilks, united by the project of investigating simplicity as a general principle of cognition. The central hypothesis driving these studies is that *our cognitive system favours simple interpretations* (mental models/hypotheses) of the data; put differently, we are wired to search for simple patterns in the world. We'll refer to this as the cognitive simplicity hypothesis (CSH).

What makes a pattern, or a hypothesis, simple? Typically, cognitive scientists employ an information-theoretic measure of simplicity (e.g. as provided by Kolmogorov complexity theory, or Shannon's information theory) in a universal coding language. The general idea is that the simplicity of a pattern can be measured by the extent to which it compresses–provides a compact encoding of–the data; the simplest pattern, corresponding to the shortest coding, provides the least redundant representation of the data.[22]

Thus far, CSH has been vindicated by a host of empirical studies from various sub-domains[23] showing that this increasingly well-documented simplicity bias supports successful explanations and predictions. From this vast literature, we single out for mention a handful of studies that focus on the role of simplicity in language learning/acquisition (Onnis et al. 2002; Hsu et al. 2013; Chater et al. 2015) and language evolution (Christiansen et al. 2006; Chater and Christiansen 2010; Culbertson and Kirby 2016), and present what we regard as their key highlights and points of contact with minimalist inquiry.

approach is implemented. To our knowledge, *this* particular slant remains under-discussed in the generativist literature (though not within the cognitive sciences, as we'll see). That said, there is also a sense in which the approach we are promoting is broader in scope than Boeckx's. Like him, we pressing for a serious collaboration between linguistics and cognitive science, *with respect to their shared subject matter*. But we are advocating just as strongly for a collaboration with philosophy of science, *with respect to the aims, methods and vehicles of linguistic inquiry* (see Sects. 5–6).

[22] The invariance theorem (Li and Vitányi 1997) ensures that the shortest description of any object is language-invariant (up to a constant).

[23] Including: concept learning (Feldman 2003); perceptual organization and category learning (Gershman and Niv 2013; Pothos and Chater 2005); function acquisition (Narain et al. 2014); causal reasoning (Lombrozo 2016; Bonawitz and Lombrozo 2012); sensorimotor learning (Genewein and Braun 2014). For a recent survey, see (Feldman 2016).

Recall the generative solution to the acquisition problem: a language faculty endowed with a rich, innate UG. This has two crucial explanatory benefits: it accounts for the universality of language, and it 'compensates' the paucity of data available to the child. The latter is a central ingredient of the so-called 'poverty of stimulus' argument for UG, which emphasises that said data is not only quantitatively limited, but also almost entirely positive, thus making the putative task of learning language from data alone implausibly hard, if not impossible. While the argument–which is cast as an instance of inference to the best explanation–continues to hold sway among generativists, recent empirical studies on language learning point to a way out of the problem of positive evidence. In a nutshell, one of their key conclusions is that in the presence of a general cognitive simplicity principle, *the input data is* sufficiently rich to ground language acquisition. The significance of this result cannot be understated, we think: if CSH continues to hold up under future empirical scrutiny, it would seem that the acquisition problem could be put to rest *without needing to postulate any innate linguistic content*.

Indeed, if the above results indicate that we can do without innate linguistic content, a second set of studies suggest that we *should* forego such assumptions. To see this, recall the minimalist strategy to address Darwin's Problem: shift content from the first to the third factor, and empty UG of any redundant content. While this is promising from a naturalistic perspective, at least insofar as it is intended to align linguistic theory with evolutionary theory, we suggest that, in light of the following, minimalists as a community can and should take their strategy one step further.

Suppose we ask: what's left in FL once any and all redundant content is stripped away from UG? Minimalist answers will vary (even significantly), but most will make reference to at least one specific *linguistic* property, or mechanism; in the terminology borrowed from the cognitive sciences, a domain-specific hard constraint. In the best-case minimalist scenario, only one such constraint would be required to explain–in conjunction with more general cognitive mechanisms–everything from language acquisition to language evolution. However, even this ideal model proves problematic from an evolutionary standpoint. The problem is that a domain-specific hard constraint, of the sort that would qualify as first-factor content, is unlikely to have evolved–even more so given the relatively recent appearance of language. On the other hand, the evolution of domain-*general*, weak constraints (or biases) seems well-supported by evolutionary theory. In particular, there seems to be mounting evidence to the effect that one such constraint is none other than the cognitive simplicity principle.

Against this backdrop, a number of recent studies have set out to investigate the conjecture that language may be the result, not of a specific evolutionary adaptation[24] but rather of the interplay of evolved, weak biases and cultural evolution. One such argument is made by Culbertson and Kirby (2016), who start off by distinguishing two ways in which a property may be specific of a given cognitive domain: the property may have evolved for a specific functional purpose, or it may have evolved for either a different or a domain-general purpose, eventually coming to interact with a specific cognitive system in a unique way. The authors then argue that language evolution[25] is most plausibly captured by the latter explanatory route. Their argument draws on two main sets of results, obtained via computational models of language evolution. The first set shows that a genetically determined universal grammar–the sort of innate content posited by generative theories–is unlikely to have evolved, either by natural selection or by other evolutionary mechanisms.[26] The second set suggests, first, that cultural evolution has an amplifying effect on weak cognitive biases; secondly, that "weak biases for language learning are *more* evolvable by virtue of cultural evolution's amplifying effect" (2016, p. 3). From the foregoing, the authors correctly draw the cautious conclusion that,

> While this does not categorically rule out the existence of very strong (or inviolable) biases that have evolved

[24] Nor of a 'miracle' or an evolutionary jump, as some minimalists have occasionally suggested. For recent discussion of a saltationist account of language evolution, see Kinsella (2009), Di Sciullo and Boeckx (2011) and Tallerman and Gibson (2012).

[25] That is, the evolution of "the linguistic system—its architecture, the representations it operates on, the constraints it is subject to" (Culbertson and Kirby 2016, p. 1).

[26] Such as the Baldwin Effect, "whereby traits that were previously acquired through experience become nativised" (Culbertson and Kirby 2016, p. 2).

> specifically for language, it clearly suggests we should not treat them as the default hypothesis. (2016, p. 9)

More interestingly still, they make a compelling case for the hypothesis that several linguistic phenomena could be domain-specific *effects* (vs. hardwired constraints) of a domain-general simplicity bias as the latter interacts with 'linguistic representations'–that is, in the terminology of the previous sections, with second-factor content, i.e. the linguistic environment (see also Thompson et al. 2016).

Where does the foregoing leave us? Earlier we noted how recent attempts to flesh out SMT have led minimalists to place more weight on the third-factor hypothesis. However, ensuing proposals have struggled to genuinely distance themselves from the dominant model of FL as a language-specific module structured by innate, language-specific hard constraints. While this is certainly understandable from a sociological perspective, it seems unsatisfactory by naturalistic standards. This becomes starkly evident once we take into account the vast array of empirical studies that point, rather convincingly, to the implausibility of said model; and which furthermore offer an alternative, scientifically robust framework within which solutions to both Plato's and Darwin's Problems appear well within reach.

In the next two sections we push for an analogous 'naturalizing' move with respect to theory-level simplicity claims. One of its main upshots will also mirror an important takeaway from the present section: namely, that serious pursuit of a justification of theory-simplicity may require breaking down inter-disciplinary barriers and–in this specific instance–looking at what philosophers have to say.

5. Theory-simplicity: a compatibilist alternative

As noted in earlier sections, and bracketing issues of object-theory conflation for the moment, generative appeals to simplicity as a theoretical virtue have sought to fall in line with a general tendency, in science and philosophy, to favour simple theories, models, explanations, etc. However, such appeals have rarely been accompanied by in-depth reflection on the questions of how theory-simplicity ought to be defined, measured, justified, and traded-off.[27] In fairness, generativists are hardly the exception in this regard; nonetheless, given the prominence ascribed to simplicity in minimalist theorising, we suggest such a reflection should be delayed no further.

To this end, a natural source of inspiration is the philosophical discussion on the role of theoretical values in scientific practice. Within this debate, analyses of so-called aesthetic values–including simplicity–traditionally fall into one of two camps: those that construe aesthetic values as 'merely' *pragmatic* criteria, and those that ascribe a more substantive, *epistemic* role to these notions.[28] Construals of the first sort typically place a strong emphasis on the variability, relativity and even subjectivity of aesthetic (and any other non-evidential) values; on this view, simplicity is cast in a strongly instrumentalist light, with connotations of 'easy to use', and the like. By contrast, accounts of the second sort regard all such values as truth-conducive–albeit to different degrees, with greater weight being allocated to evidential criteria such as empirical adequacy and predictive power.[29]

Here we sketch a compatibilist alternative to the above, that draws on recent proposals according to which aesthetic values do indeed serve a substantive epistemic function in scientific practice, without however relinquishing their pragmatic connotation (Breitenbach 2013; de Regt 2017; Kosso 2002; Ivanova 2017). More specifically, on this view aesthetic values are epistemically 'active' insofar as they are indicative of, and conducive to, *understanding* (of relevant target phenomena), where the latter is a central aim of science. Our main contention is that an analogous recalibration of the aims of inquiry would be both recommendable and potentially fruitful in the generative

[27] In what follows, unless otherwise indicated any mention of simplicity should be understood to refer to theory-simplicity.

[28] See e.g. Baker (2003), Barnes (2000), Schindler (2018) and Van Fraassen (1980).

[29] To repeat, generativists have very rarely engaged with this debate; to our knowledge, the only two exceptions are Barrios (2016) and Ludlow (2011), more about which in Sect. 5.3.

context. Given that this move hinges in turn on the epistemic notion of understanding, some stage-setting is appropriate at this point. We give a brief overview of the ongoing philosophical debate surrounding the notion of understanding as an aim of science in Sect. 5.1, which we then tie in with discussions of simplicity in 5.2. Against the resulting backdrop, we then comment on two extant analyses of simplicity in the context of generative linguistics (Sect. 5.3). Ultimately, we'll see that neither is entirely satisfying precisely because they fail to distance themselves from the traditional adversarial narratives of science as either a truth-bound enterprise, or as subject to mere empirical adequacy standards. Building on the foregoing, Sect. 6 outlines what the proposed methodological and philosophical shift might 'look like' in generative linguistics.

5.1. Scientific understanding: a brief overview of the debate

Understanding is currently (and has been for the past decade or two) a hot topic both within general epistemology and in the philosophy of science. While the landscape of this philosophical debate is heterogeneous with respect to what we might label 'local' issues (more about which shortly), it is fair to say that there is a broad consensus according to which understanding is a cognitive-epistemic achievement which is (i) *more demanding than knowledge*; and (ii) tightly enmeshed (if not identical) with the central scientific aim of *producing explanations* of natural and social phenomena.

Both (i) and (ii) are fairly nebulous as they stand, of course. Unsurprisingly, disagreements have arisen wherever attempts to sharpen either thesis have been made. This is true more pointedly of (i): here, the issue that has proven to be particularly divisive concerns the relation of understanding to knowledge. More specifically, the main sticking point is whether understanding is a subspecies of knowledge, or if the two are entirely distinct epistemic achievements (see e.g. Grimm et al. (2017) for an excellent introduction). Within the former camp, moreover, further disagreement concerns whether understanding and knowledge share some vs. all of the same satisfaction conditions (minimally: truth, belief, justification).[30]

In what sense, then, is there any kind of agreement over (i)? The consensus is that understanding requires 'something extra' over and above knowledge: namely, it requires that the subject grasp (at least some among) the salient explanatory connections within the domain that is the target of understanding. For instance, Kvanvig writes that

> One can know many unrelated pieces of information, but understanding is achieved only when informational items are pieced together by the subject [...]. [This is the] crucial difference between knowledge and understanding: that understanding requires, and knowledge does not, an internal grasping or appreciation of how the various elements in a body of information are related to each other in terms of explanatory, logical, probabilistic, and other kinds of relations [...]. (2003, p. 192f.)

More generally, different authors have offered slightly different characterisations of the notion of grasping.[31] By and large however it is agreed that grasping is not reducible to propositional attitudes such as knowledge or belief. We follow Bailer-Jones (1997) and Reutlinger et al. (2017) (who in turn seem to express an implicit consensus in the literature) in allowing that grasping is philosophically primitive, though not scientifically so. Thus, insofar as it is a cognitive activity, grasping is a legitimate object of study for the cognitive sciences; but philosophically, it seems perfectly acceptable to have the buck stop here.[32] Importantly,

[30] For instance, Grimm and Khalifa take understanding to be factive, citing different arguments for this claim (Grimm 2006; Khalifa 2013). By contrast Elgin and Zagzebski side with the non-factivist camp (Elgin 2017; Zagzebski 2001). Similarly, on whether understanding requires at least one among belief, justification, anti-luck conditions see for example Grimm (2006), Kvanvig (2009) and Dellsén (2017).

[31] For instance, some push for a conception of grasping as a sort of ability, e.g. the ability to manipulate the relations between propositions. See among others Hills (2016).

[32] Only one attempt has been made to further analyse the concept of grasping, that we know of, by Janvid (2018).

grasping is acknowledged to be independent of truth, even as the (non-) factivity of understanding continues to be hotly debated. This issue is more salient than others, in the context of our discussion, because it speaks to the question of whether only true (or probably or approximately true) scientific theories are to be considered reliable vehicles of scientific explanation and therefore understanding, or whether other kinds of vehicles might be included in this class. This takes us back to item (ii).

The central questions here are, first, what counts as an explanation–of the sort produced by scientists in their effort to advance their (individual and/or collective) understanding of the world. The second question is whether understanding can be mediated by different epistemic vehicles (beyond theories in the traditional, propositional sense) or is instead restricted to a specific subclass of such vehicles. On both these counts, the literature offers a picture that is more distinctively pluralistic than divisive. Thus, more or less peacefully co-existing in the current landscape are those who argue that understanding can be yielded by causal, how-actually explanations (Khalifa 2017); non-causal, how-actually explanations (Lipton 2009); how-possibly explanations (Reutlinger et al. 2017); successful classifications (Gijsbers 2013); non-propositional representations (de Regt 2017); models and idealizations (Elgin 2007; Strevens 2016); and perhaps fictions and more besides (Lawler 2019).

5.2. Theoretical values and scientific understanding

What does simplicity have to do with the foregoing? The consensus view that emerges from the literature is that aesthetic values, alongside more 'canonical' values such as consistency or predictive power, play an often crucial role in the subject's achievement of understanding of the target via one or more relevant epistemic vehicles. Crucially, they contribute to this epistemic goal precisely in virtue of their pragmatic dimension.

One way to flesh out this idea is via de Regt's notion of intelligibility of scientific theory. In line with the above-mentioned literature, de Regt (2009, 2017) identifies as a central aim of science what he calls 'understanding a phenomenon', or UP: the understanding that is provided by having an adequate explanation of the phenomena being investigated.[33]

Understanding (i.e. UP) is thus a relation between subject and world; crucially for de Regt, it is mediated by intelligible theories, where intelligibility is defined as

> the value that scientists attribute to the cluster of qualities of a theory (in one or more of its representations) that facilitate the use of the theory.[34] (2017, p. 40)

Notice that while UP has a distinctively epistemic ring to it, intelligibility has expressly pragmatic overtones. De Regt's key thesis is that the latter is a necessary condition for the former: that is, successful explanations of phenomena require intelligible theories. Therefore, since theoretical values help shape intelligible theories, they are themselves preconditions of explanatory understanding.

By explicitly recognising that the epistemic and the pragmatic dimensions are thus enmeshed, the perspective developed by de Regt and others is a dynamical one, certainly compared to more established, incompatibilist construals. Indeed, a distinctive and shared feature of the former is the importance ascribed to context in shaping UP, by acknowledging the variability of theoretical values and their respective weights along multiple dimensions: through history, across domains of inquiry, between scientific communities; and among members of these

[33] The notion of explanation that de Regt has in mind is more flexible than traditional—for instance, strictly causal—conceptions. In other words, de Regt is among those who subscribe to a pluralist view of explanation (and of epistemic vehicles); in particular, on his view "all explanations are, in a broad sense, arguments. An explanation is an attempt [...] to provide understanding of the phenomenon or the situation by presenting a systematic line of reasoning that connects it with other accepted items of knowledge" (de Regt 2017, p. 25). We are sympathetic to de Regt's conception, above all because we are sympathetic to its pluralist spirit.

[34] Two things should be noted here. The first is that de Regt's discussion concerns the broader class of theoretical values, including but not limited to aesthetic values; for instance, he notes that "Causal structure is a quality that is often regarded as enhancing the intelligibility of theories" (de Regt 2017, p. 109). Secondly, as with the notion of explanation, de Regt favours an interpretation of 'theory' that is loosened to encompass also models, idealizations, experimentations, etc.

communities, depending on "background knowledge, metaphysical commitments, and the virtues of already entrenched theories" (de Regt 2009, p. 31).[35] Crucially, this multifaceted context-sensitivity doesn't collapse into relativism: as Douglas (2013, p. 802) puts it, "the proof will be in the pudding [...], and the pudding is relatively straightforward to assess. [...] We will be able to tell readily if the instantiation of a pragmatic-based value in fact proves its worth."

One of the many merits of de Regt's account is that it pays the history of science its due attention, offering detailed case studies (mainly from the history of physics) as a means both to illustrate his proposal, and to ensure it remains tethered to scientific practice.[36] However, while de Regt makes a compelling case for a robustly contextualist account of theoretical values, we find that he ends up obscuring a particularly interesting fact as a result: namely, that while many theoretical values have come and gone over the course of the history of science (e.g. visualizability), the cluster of so-called aesthetic values has remained a more or less stable fixture throughout. This observation is one of the premises of Breitenbach's account, to which we now turn.

Like de Regt, Breitenbach argues that understanding is a ternary relation between theory, world and scientist; more specifically–with an emphasis that sets her apart from de Regt–*the scientist's cognitive structure and capacities*. Following the declared Kantian inspiration of her account, Breitenbach construes aesthetic judgments in science as second-order responses to "our awareness of the suitability of our

[35] Forster and Sober (1994) also argue for a local justification of simplicity, from rather different premises. Their argument, very much boiled down, runs as follows: the main goal of model selection is predictive accuracy (rather than probable truth); insofar as simplicity minimises the risk of overfitting the data, it also favours predictive accuracy; therefore, simplicity should be favoured—in the context of model selection problems. See also (Sober 2002).

[36] Three of the final chapters of (de Regt 2017) are devoted to the discussion of, respectively: the intelligibility of Newton's gravitation theory; the role of mechanical models as vehicles of understanding in 19th century physics; the role of visualization—as a criterion of intelligibility and therefore a condition of explanatory understanding—in the transition from classical to quantum physics (in particular, its role in the Heisenberg/Schrödinger debate over the superiority of matrix mechanics versus wave mechanics, respectively, as means to understand atomic phenomena).

intellectual capacities for making sense of the world around us" (2013, p. 92). Importantly, aesthetic judgments are thus neither directly about the world, nor about the theory per se. Rather, they are "essentially self-reflective," in that they reveal–mark our awareness of–the attainment of a certain harmony between our cognitive makeup and the world, mediated by our representations (theories, models) of the latter. Therefore, *aesthetic values are conditions of understanding*. Moreover, insofar as this is the case we are also *justified in pursuing simplicity*, unity, beauty etc. in our theories: for, while it is neither necessarily nor contingently true that simple theories will provide understanding (much less be truth-conducive), nonetheless they

> condition the possibility of such understanding, [and] providing such understanding is an essential requirement for any successful theory. (Breitenbach 2013, p. 96)

Together, Breitenbach's and de Regt's proposals offer a powerful and compelling account of the role of aesthetic values, including simplicity, in shaping scientific practice. Moreover, as we'll see in Sect. 6, the conception of scientific practice (specifically, its aims and methods) underlying these and similar accounts offers a novel and fruitful vantage point from which to re-examine linguistic practice.

5.3. Barrios and Ludlow on simplicity in generative linguistics

To complete our stage-setting operation we now examine two separate discussions of theory-simplicity in the philosophy of linguistics offered by, respectively, Barrios (2016) and Ludlow (2011) (see footnote 29). In so doing we hope to further elucidate the merits of our preferred, alternative construal of this notion. The first thing to note is that both Barrios's and Ludlow's analyses are to a certain extent entirely compatible with, in particular, de Regt's account of simplicity (among other aesthetic values). In particular, both authors agree that ascriptions of theory-simplicity are sensitive to contextual factors, in the sense that they vary from one scientific community to another, between stages of inquiry and scientific periods, and over time.[37]

However, whereas Barrios correctly recognises and indeed emphasises the varied epistemic roles played by simplicity considerations *vis-à-vis* the explanatory aims of science, Ludlow strongly downplays (indeed, ignores) the connection between the pragmatic character of simplicity and the epistemic function it serves in contexts of theory construction, choice etc. Thus, Ludlow argues that simplicity, as this notion applies to scientific theories (as opposed to subject matter) in general, and linguistic theories in particular, is nothing more than a pragmatic criterion, narrowly construed as synonymous with 'easy to use': "when we look at other sciences, in nearly every case, the best theory is arguably not the one that reduces the number of components from four to three, but rather the theory that allows for the simplest calculations and greatest ease of use" (Ludlow 2011, p. 158).[38]

Despite the above-mentioned overlap with the contextualist theses propounded by de Regt, Ludlow's argument for this 'ease of use' thesis is unconvincing, we find. This is in large part because it rests on a false dichotomy: namely, that simplicity must be conceived of either as an objective, "absolute" and universal property of theories (possibly complemented by a realist metaphysical justification about the simplicity of reality); or as an always subjective, relative, strictly pragmatic connotation of those theories that allow us to "accomplish our goals with the minimal amount of cognitive labor" (2011, p. 152).

[37] Cf. for instance Ludlow's Theses I–III (2011, pp. 161–162).

[38] Ludlow very briefly acknowledges that alongside theory-simplicity, MP is also motivated by a second notion whose role is essentially that of an explanatory goal: namely, to reduce the subject matter of linguistics to one that is more fundamental ("low level biophysical processes" (Ludlow 2011, p. 160); but cf. footnote 15). This is of course the interpretation of simplicity underlying the third-factor hypothesis, which we discussed in Sect. 4—albeit not in terms of reduction. In large part, this is because the term 'reduction' is very rarely employed by generativists, who have indeed occasionally explicitly rejected this interpretation of their practice. But we do *de facto* discuss reduction (albeit horizontal—to domain-general cognitive principles—rather than vertical—to low-level processes) in the context of discussing the third-factor hypothesis (that is, where simplicity becomes an explanatory goal).

In a sense, we might charitably say that Ludlow's account stops short at de Regt's intelligibility condition; indeed, on the few occasions in which Ludlow mentions understanding (e.g.: "the clearest sense we can make of [simplicity] is [...] in terms of 'simple to use and understand' " (2011, p. 152)) it is reasonably clear that he has in mind what de Regt terms 'understanding a theory.' The merit of the latter's account is that it explores the connection between such pragmatic considerations and the wider explanatory aims and achievements of science. By contrast, as noted above Barrios does acknowledge such connections, both with respect to linguistic inquiry and to science at large. For instance, Barrios offers a reconstruction of generative history which–not unlike the reconstruction presented in our Sect. 2–emphasises the parallelism between the changing role of simplicity on the one hand, and the goals of linguistic inquiry (observational adequacy, descriptive adequacy, explanatory adequacy, explanatory depth) on the other; he also offers an orthogonal analysis that identifies some of the traditional interpretations of simplicity (unification, parsimony) as underlying specific stages of linguistic theory.

Without entering into a detailed discussion of Barrios's rich analysis of simplicity throughout generative history–much of which we agree with–here we merely comment on the main difference between that proposal and the present one. In a nutshell, the divergence stems from our respective conceptions of the aims of scientific (and linguistic) inquiry, as well as of the methods deployed to achieve such aims. As to the former, Barrios seems on the whole to side with a more orthodox conception according to which science (and therefore linguistics) aims at the truth, or some reasonably close proxy. Similarly, Barrios entertains a more or less traditional conception of the vehicles of scientific inquiry, that construes the latter class as exhausted by theories in the standard sense. In contrast, the proposals we are aligning ourselves with support a conception of scientific vehicle that is both more flexible–the relevance of which will become clearer in Sect. 6–and (therefore) more faithful to actual scientific practice. In sum, in both these respects we part ways with Barrios over much the same concerns that separate current accounts of scientific understanding from the more traditional analyses of this notion.

We submit that the perspectives on theory-simplicity presented in this section have potentially significant repercussions for linguistic inquiry. In the next section we finally put the pieces together, and sketch what we see as a promising research agenda for generative linguistics, philosophy and cognitive science.

5.2. Theoretical values and scientific understanding

Up until now, we have discussed language acquisition and evolution as largely separate problems. But the two share an important connection, insofar as their respective generative solutions pull in opposite directions: acquisition requires rich, innate linguistic content, and evolution requires a thin, deflated UG. This tension is defused, however, in light of the proposal sketched in Sect. 4: that is, if we set aside the idea that '*something* must be special to language', and countenance the hypothesis that language acquisition could be explained in terms of second- and third-factor content alone. Indeed, we maintain this would qualify as an appealing approach *by minimalist standards*, for several reasons: (1) current empirical research suggests that any 'solution' to Plato's Problem would feature simplicity (as a general cognitive principle) among its main explanatory factors; (2) the hypothesis of a cognitive simplicity principle seems to breathe new life into the early generative insight (Chomsky 1965) that some sort of internal simplicity criterion participates in language acquisition;[39] (3) by subsuming language acquisition under a broader cognitive account, (a) the resulting explanation would meet several theoretical desiderata such as coherence, unification and, of course, simplicity; (b) the account would also meet both kinds of naturalist standards–ours, and the minimalist's (cf. Sect. 3). These reasons are further compounded by a fourth: namely, that the integration of minimalist inquiry into cognitive science would allow for a *unified* treatment of both Plato's and Darwin's Problems.

To reiterate, we think that while the foregoing does require a perspective shift on the minimalist's part, it can still be reconciled with

[39] See also Yang (2017).

the spirit of (at least some) minimalist tenets. At the beginning of Sect. 3, we remarked on the fluctuations in the interpretation of (both object- and theory-) simplicity between and even within competing frameworks. In fact, diachronic analyses such as ours reveal a subtler trend than this, especially where object-simplicity is concerned. That is, over and above any and all local variations, what remains fixed is the idea that *object-simplicity is language-specific*. Our proposal would require this idea to be revisited rather than abandoned: specifically, to shift from thinking of FL as intrinsically simple (perhaps as a corollary of a sweeping generalisation about the simplicity of nature), to thinking that FL *inherits its simplicity from domain-general features of our cognitive system*.

Indeed, we're making a broadly parallel point about theory-simplicity. What transpired from Sects. 2–3 is that as a result of their commitment to a hard-nosed realism combined with the Galilean style, minimalists have come to hold an unnecessarily narrow perspective on the available 'meta'-explanatory options. Among other things, this means that truth (or approximate truth, representational accuracy, etc.) stands unchallenged as the do or die of any one account, at the expense of other epistemic benefits. Here, too, our proposal is of a hermeneutic rather than revolutionary stripe. We're not suggesting that minimalists toss out any (much less all) of the theoretical achievements accrued so far. In fact, we're urging that minimalists themselves avoid doing so: rather than holding theoretical products to a single uncompromising standard of truth, other explanatory and epistemic benefits, sanctioned by successful sciences, should be considered.

In addition, it seems to us that the foregoing dovetails very nicely with the philosophical analyses of the role of aesthetic values described in Sect. 5.2. On the one hand de Regt's contextualist account offers an illuminating interpretative key on the fluctuating conceptualisation of simplicity in the course of generative history. Furthermore, both de Regt and yet more explicitly Breitenbach ascribe a more prominent role to theoretical values–including simplicity–in scientific practice, as a result of carving out the relation between scientist-theory-world in a novel way. A third point of contact is seen most clearly by noting a salient difference between the two accounts: while de Regt's main concern is to elucidate the ways in which theoretical values contribute to scientific understanding, Breitenbach is more interested in where these values 'come from'. And, once her proposal is stripped of its Kantian overtones, what remains is a cognitive hypothesis: namely, that aesthetic judgments are the result of the subject's cognitive makeup, and of the interaction between the latter and the world, via theory.

In light of these observations, a few interesting projects suggest themselves. First, we think it would be a fruitful minimalist exercise to examine past and current linguistic practice by the lights of the above philosophical accounts. There are many ways one could implement this somewhat vague suggestion. In what follows we sketch just one of these.

In Sect. 5, we made a point of emphasizing the pluralist orientation of the debate on understanding; this is witnessed, for instance, by the gradual broadening of accepted construals of the notion of explanation, to encompass even mutually incompatible conceptions. Of particular interest is the manifestation of such pluralist tendencies with respect to the vehicles of scientific understanding. We've seen this to be a varied class (Sect. 5.1); even more so when we take into account the heterogeneity of its proper subclasses. Indeed the single most diverse of these subclasses is also the most resourced by working scientists: namely the class of scientific models, minimally construed as (more or less idealized) representations of a target phenomenon. That models come in many shapes and forms is well known; for instance, two models about a same target phenomenon P may differ in terms of the degree of abstraction incorporated in their respective representations of P. Models can be highly realistic and concrete (e.g. scale models) or highly idealized and abstract (e.g. toy models). Most interestingly for our purposes, even models that sit at the latter end of the spectrum–that is, even models that are highly simple, idealized and literally false of their target, known as toy models–are widely recognized to be vehicles of scientific understanding.[40]

[40] A well known example is the Schelling model of racial segregation (Weisberg 2013). The model's target is the phenomenon of segregation in urban areas; its representation of this phenomenon is highly simplified (in that it makes only very few assumptions about the target) and idealized (in that its main assumptions contain deliberate distortions, such as the absence of difference-making socioeconomic factors). Since its inception, Schelling's model (also known as the checkerboard model) has been widely used by social scientists as well as coopted by philosophers to study and illuminate previously undetected features of segregation (and segregation-like) phenomena.

In what way do models so far removed from reality produce, or advance, understanding of their target phenomena? The widely accepted answer is that they do so *precisely as a result of* their deliberate suspension and/or distortion of explanatory factors. More generally, it is (also) in virtue of their extreme simplicity that toy models throw light on phenomena that are either too complex to study directly, or where it is still unclear which factors are genuinely explanatory, and so on. Thus, even toy models are qualified to deliver understanding: specifically, as argued for instance by Reutlinger et al. (2017), they (can) provide a *potential explanation* of their target phenomenon, as a result of which they (can) produce or enhance *how-possibly* understanding of the phenomenon in question.

We think that the foregoing–and more generally, the broader debate on ways in which different epistemic vehicles can function as gateways to scientific understanding–could lead to powerful new insights within generative practice; conversely, we think that generative linguistics should be included in the philosophical conversation on the aims and methods of science. In order to implement this idea, a first and prerequisite step must be for the generative community to liberalize their extant conception of epistemic vehicle, in particular to encompass those which do not satisfy a strict factivity clause (e.g. idealized models). A subsequent key step would then be to reinterpret specific generative theories and hypotheses–attributing ever-increasing simplicity to FL–as candidate vehicles of one or more kinds of understanding.

As a prime illustration, consider P&P. As we saw in Sect. 2, P&P retained a lasting influence (up to and including the early years of MP) insofar as it offered a simple and attractive answer to Plato's Problem, in terms of a relatively small number of abstract, universal, innate principles together with parameters that are switched on or off in response to environmental linguistic stimuli. What went wrong? The standard answer is that P&P is incompatible with evolutionary theory. But another way of seeing things is that P&P was judged (and therefore eventually discarded) *qua* purportedly veridical theory. However, once we liberalize the working conception of epistemic vehicle, new options open up. In particular, it becomes very natural to reinterpret P&P as a highly idealized model of FL: one that suspends at least one explanatory factor (the acquisition process, which is relegated to an infallible on/off switch) and distorts others (the bulk of the explanatory burden is borne by innate, domain-specific content). Once these substantial idealizations are acknowledged, it becomes quite clear that P&P, while implausible and indeed unviable as a veridical theory, can however yield understanding in the form of a potential explanation of its target phenomenon. Thus, P&P helps shed light on questions such as: How much of the explanatory burden of language acquisition can be pushed onto innate, language-specific content? And: Which among the acknowledged explanatory factors (innate linguistic content, acquisition process, primary linguistic data) are genuine difference-makers? And so on. An immediate upshot is then that P&P needn't be discarded *just because* it is false of the actual world. It should rather be judged on its merits *as a vehicle of understanding* of language acquisition.[41]

In closing, we mention just two more promising angles of future inquiry. First, we think that generative debates hold deep philosophical interest, whereas they have been largely ignored by mainstream philosophy. In particular, we hope to have shown that generative linguistics makes for an intriguing case study on the relation between criteria of scientific understanding, explanatory adequacy, and different interpretations of simplicity.[42]

Finally, it would be an interesting project to examine Breitenbach's hypothesis itself from an empirical perspective, and more specifically to investigate (i) the cognitive underpinnings of understanding, and (ii) the connection between the latter and the cognitive simplicity principle.

[41] Indeed, we think a case can be made to the effect that P&P yields understanding not just of a modal variety—as suggested here—but also of both heuristic and pedagogical ones (cf. Reutlinger et al. 2017). We are developing both ideas in preparation for a separate article.

[42] E.g. ontological versus syntactic (Baker 2016); anti-quantitative vs. anti-superfluity (Barnes 2000); agnostic vs. atheistic Ockham's razor (Sober 2015); Ockham's razor vs. Ockham's laser (Baron and Tallant 2018); quantitative vs. qualitative parsimony in science and philosophy (Lewis 1973; Jansson and Tallant 2016).

7. Conclusion

This paper started with the observation that, given the centrality of simplicity in their most recent research program, minimalists ought to address the issues of justification and convergence as a matter of urgency. We then outlined and defended a naturalistic approach to both questions; crucially, the proposals outlined in Sects. 4 and 5–6 are accompanied by robust *justifications* of, respectively, the hypothesis that simplicity is a property of FL (insofar as it is a general cognitive principle that interacts with FL to produce domain-specific effects) and the adoption of simplicity as a theoretical value (insofar as simplicity, along with other aesthetic values, is conducive to understanding).

Just as importantly, the proposed account offers a sharper and more nuanced characterisation of both object- and theory-simplicity that rules out the possibility of further conflation of these notions. Conversely, with these sharpened notions in hand it becomes possible to rigorously assess the minimalist expectation that the two should converge.

Finally, we hope to have shown that embarking on a genuinely collaborative path promises to be a fruitful endeavour for minimalists, philosophers and cognitive scientists alike.

References

Al-Mutairi, F. R. (2014). *The minimalist program: The nature and plausibility of Chomsky's biolinguistics* (Vol. 143). CUP.

Ankeny, R., Chang, H., Boumans, M., & Boon, M. (2011). Introduction: Philosophy of science in practice. *European Journal for Philosophy of Science*, *1*(3), 303.

Bailer-Jones, D. (1997). *Scientific models: A cognitive approach with an application in astrophysics*. PhD thesis, University of Cambridge.

Baker, A. (2003). Quantitative parsimony and explanatory power. *The British Journal for the Philosophy of Science*, *54*(2), 245–259.

Baker, A. (2016). Simplicity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (winter ed.). Stanford: Metaphysics Research Lab, Stanford University.

Barnes, E. C. (2000). Ockham's razor and the anti-superfluity principle. *Erkenntnis*, *53*(3), 353–374. Baron, S., & Tallant, J. (2018). Do not revise Ockham's razor without necessity. *Philosophy and Phenomenological Research*, *96*(3), 596–619.

Barrios, E. (2016). Simple is not easy. *Synthese*, *193*(7), 2261–2305.

Boeckx, C. (2006). *Linguistic minimalism: Origins, concepts, methods, and aims*. Oxford: OUP.

Boeckx, C. (2009). The nature of merge: Consequences for language, mind, and biology. In M. Piattelli-Palmarini, J. Uriagereka, & P. Salaburu (Eds.), *Of minds and language: A dialogue with Noam Chomsky in the Basque Country* (pp. 44–57). Oxford: OUP.

Boeckx, C. (2010). Linguistic minimalism. In B. Heine & H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (pp. 485–505). Oxford: OUP.

Boeckx, C. (2014a). *Elementary Syntactic Structures: Prospects of a Feature-Free Syntax*. Cambridge Studies in Linguistics. Cambridge University Press.

Boeckx, C. (2014b). What principles and parameters got wrong. In M. C. Picallo (Ed.), *Linguistic Variation in the Minimalist Framework*. Oxford: OUP.

Boeckx, C. (2016). Considerations pertaining to the nature of logodiversity. In *Rethinking parameters* (pp. 64–104). Oxford University Press.

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*, *48*(4), 1156.

Breitenbach, A. (2013). Aesthetics in science: A Kantian proposal. In *Proceedings of the Aristotelian Society* (Vol. 113, pp. 83–100). Wiley Online Library.

Chater, N., & Christiansen, M. H. (2010). Language evolution as cultural evolution: How language is shaped by the brain. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 623–628.

Chater, N., Clark, A., Goldsmith, J. A., & Perfors, A. (2015). *Empiricism and Language Learnability*.

Oxford: OUP.

Chomsky, N. (1951). Morphophonemics of Modern Hebrew. MA thesis, University of Pennsylvania, New York.

Chomsky, N. (1957). *Syntactic Structure*. Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1975/1955). *The logical structure of linguistic theory*. University of Chicago Press.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Kluwer.

Chomsky, N. (1993). *Lectures on government and binding: The Pisa lectures*. Number 9. Walter de Gruyter.

Chomsky, N. (1995). *The Minimalist Program*. Current Studies in Linguistics 28.

Chomsky, N. (1996) *Powers and Prospects: Reflections on Human Nature and the Social Order*. Pluto Press.

Chomsky, N. (2004). *The generative enterprise revisited: Discussions with Riny Huybregts, Henk van Riemsdijk, Naoki Fukui and Mihoko Zushi*. Berlin: Mouton de Gruyter.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, *36*(1), 1–22.

Chomsky, N. (2007). Approaching UG from below. In U. Sauerland & H.-M. Gärtner (Eds.), *Interfaces + recursion= language? Chomsky's Minimalism and the View from Syntax-Semantics* (pp. 1–29). Berlin: Mouton de Gruyter.

Chomsky, N. (2009). *Cartesian linguistics: A chapter in the history of rationalist thought*. Cambridge: CUP.

Christiansen, M. H., Reali, F., & Chater, N. (2006). The Baldwin effect works for functional, but not arbitrary, features of language. In *The Evolution of Language* (pp. 27–34). World Scientific.

Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, *6*, 1964.

de Regt, H. (2009). Intelligibility and scientific understanding. In H. de Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific Understanding: Philosophical Perspectives*. Pittsburgh: University of Pittsburgh Press.

de Regt, H. W. (2017). *Understanding scientific understanding*. Oxford: OUP.

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

Di Sciullo, A. M., & Boeckx, C. (2011). *The biolinguistic enterprise: New perspectives on the evolution and nature of the human language faculty* (Vol. 1). Oxford: OUP.

Douglas, H. (2013). The value of cognitive values. *Philosophy of Science*, *80*(5), 796–806. Elgin, C. (2007). Understanding and the facts. *Philosophical Studies*, *132*, 33–42.

Elgin, C. Z. (2017). Exemplification in understanding. In Grimm, S., Baumberger, C., Ammon, S. (Eds.) *Explaining understanding: New perspectives from epistemology and philosophy of science*. New York: Routledge.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, *12*(6), 227–232.

Feldman, J. (2016). The simplicity principle in perception and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(5), 330–340.

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, *45*(1), 1–35.

Freidin, R., & Lasnik, H. (2011). Some roots of minimalism in generative grammar. In Boeckx, C. (Ed.) *The Oxford handbook of linguistic minimalism* (pp. 1–26).

Genewein, T., & Braun, D. A. (2014). Occam's razor in sensorimotor learning. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1783), 2013–2952.

Gershman, S., & Niv, Y. (2013). Perceptual estimation obeys Occam's razor. *Frontiers in Psychology*, *4*, 623.

Gijsbers, V. (2013). Understanding, explanation, and unification. *Studies in History and Philosophy of Science*, *44*, 516–522.

Grimm, S. R. (2006). Is understanding a species of knowledge? *The British Journal for the Philosophy of Science*, *57*(3), 515–535.

Grimm, S. R., Baumberger, C., & Ammon, S. (2017). *Explaining understanding: New perspectives from epistemology and philosophy of science*. Abingdon: Routledge.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579.

Hills, A. (2016). Understanding why. *Noûs*, *49*(2), 661–688.

Hornstein, N., Nunes, J., & Grohmann, K. K. (2005). *Understanding minimalism*. Cambridge: CUP.

Hsu, A. S., Chater, N., & Vitányi, P. (2013). Language learning from positive evidence, reconsidered: A simplicity-based approach. *Topics in Cognitive Science*, *5*(1), 35–55.

Ivanova, M. (2017). Aesthetic values in science. *Philosophy Compass*, *12*(10), 1.

Jansson, L., & Tallant, J. (2016). Quantitative parsimony: Probably for the better. *The British Journal for the Philosophy of Science*, *68*(3), 781–803.

Janvid, M. (2018). Getting a grasp of the grasping involved in understanding. *Acta Analytica*, *33*(3), 371–383.

Kertész, A. (2010). From 'scientific revolution' to 'unscientific revolution': an analysis of approaches to the history of generative linguistics. *Language Sciences*, *32*(5), 507–527.

Khalifa, K. (2013). Is understanding explanatory or objectual? *Synthese*, *190*(6), 1153–1171.

Khalifa, K. (2017). *Understanding, explanation, and scientific knowledge*. Cambridge: CUP.

Kinsella, A. R. (2009). *Language evolution and syntactic theory* (Vol. 1). Cambridge: Cambridge University Press.

Kitcher, P. (2019). So ?who is your audience? *European Journal for Philosophy of Science*, *9*(1), 1–15.

Kosso, P. (2002). The omniscienter: Beauty and scientific understanding. *International Studies in the Philosophy of Science*, *16*(1), 39–48.

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

Kvanvig, J. (2009). The value of understanding. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Epistemic value* (pp. 95–112). Oxford: Oxford University Press.

Lappin, S., Levine, R. D., & Johnson, D. E. (2000). The structure of unscientific revolutions. *Natural Language & Linguistic Theory* (pp. 665–671).

Lawler, I. (2019). Scientific understanding and felicitous legitimate falsehoods. *Synthese* (pp. 1–29).

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Li, M., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. Berlin: Springer.

Lipton, P. (2009). *Understanding without explanation*. In *Scientific Understanding: Philosophical Perspectives*. University of Pittsburgh Press.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*(10), 748–759.

Ludlow, P. (2011). *The Philosophy of Generative Linguistics*. Oxford: OUP.

Narain, D., Smeets, J. B. J., Mamassian, P., Brenner, E., & van Beers, R. J. (2014). Structure learning and the Occam's razor principle: a new view of human function acquisition. *Frontiers in Computational Neuroscience*, *8*, 121.

Nersessian, N. J., (ed.) (1987). *The Process of Science: Contemporary Philosophical Approaches to Understanding Scientific Practice*. Number 3 in Science and Philosophy. Kluwer Academic Publishers.

Onnis, L., Roberts, M., & Chater, N. (2002). Simplicity: A cure for overgeneralizations in language acquisition? In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 24).

Pothos, E. M., & Chater, N. (2005). Unsupervised categorization and category learning. *The Quarterly Journal of Experimental Psychology*, *58*(4), 733–752.

Reutlinger, A., Hangleiter, D., & Hartmann, S. (2017). Understanding (with) toy models. *The British Journal for the Philosophy of Science*, *69*(4), 1069–1099.

Schindler, S. (2018). *Theoretical Virtues in Science: Uncovering Reality Through Theory*. Cambridge: CUP.

Smith, N., & Allott, N. (2016). *Chomsky: Ideas and ideals*. Cambridge: CUP.

Sober, E. (1975). *Simplicity*. Oxford: Clarendon Press.

Sober, E. (1978). Computability and cognition. *Synthese*, *39*(3), 383–399.

Sober, E. (2002). What is the problem of simplicity? In H. A. Keuzenkamp, M. McAlleer, & A. Zellner (Eds.), *Simplicity, inference and modelling*. Cambridge: CUP.

Sober, E. (2015). *Ockham's razors: a user's manual*. Cambridge: CUP.

Soler, L., Zwart, S., Lynch, M., & Israel-Jost, V., (eds.) (2014). *Science after the Practice Turn in the Philosophy, History, and Social Studies of Science*. Routledge Studies in the Philosophy of Science 14. Routledge.

Strevens, M. (2016). How idealizations provide understanding. In *Explaining understanding: New perspectives from epistemology and philosophy of science*. Routledge.

Tallerman, M., & Gibson, K. R., (eds.) (2012). *The Oxford handbook of language evolution*. Oxford: OUP.

Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(16), 4530–4535.

Van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: OUP.

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

Yang, C. (2017). Rage against the machine: Evaluation metrics in the 21st century. *Language Acquisition*, *24*(2), 100–125.

Zagzebski, L. (2001). Recovering understanding. In M. Steup (Ed.), *Knowledge, truth, and duty: Essays on epistemic justification, responsibility, and virtue*. Oxford: Oxford University Press.

II.
Ethics and Political Philosophy

Financing the Many Worlds: Pedagogies of (Il)liquidity

Erik Bordeleau

Bordeleau, E. (2022). Financing the many worlds: Pedagogies of (il)liquidity.
In M. Baravalle, E. Braga, & G. Riccio (Eds.), *Art for UBI (Manifesto)* (pp. 110-127).
B.r.u.n.o editore.

Who renders whom capable of what, and at what price, born by whom?
Donna Haraway, *Staying with the Trouble*

Us the living, we are a minority. A provisional minority.
Jorge Luis Borges

The implementation of wealth redistribution schemes such as a Universal Basic Income (UBI) poses a series of complex political and design challenges. But one thing is for sure: the lack of resources is not one of them. There is actually plenty for everyone – if only we manage to properly articulate collective claims on the already circulating and vastly unjustly accumulated wealth.

Capitalism is an injustice-compounding machine that must be reprogrammed. But it seems like our traditional battery of political concepts and antagonistic practices isn't quite allowing us to raise up to the challenge. For Robert Meister, a political philosopher who is teaching in the department of History of Consciousness at USC Santa Cruz alongside Donna Haraway and Anna Tsing among other illustrious colleagues, and who is the author of the seminal *Justice Is an Option: A Democratic Theory of Finance for the 21st Century* (University of Chicago Press, 2020), the operational starting point to reprogram capitalism is to be encountered at the very heart of the techno-social machinery by which current wealth is preserved and accumulated, that is: finance. To be truly effective, argues Meister, any conceivable remedy for historical injustice must also be expressible, at least initially, *in the language that the financial sector uses to value its own abundance today.* « Justice can be made more present, more embodied, more actionable in the temporalities of the present », writes Meister as he evokes his fruitful collaboration with Randy Martin, « if it is recaptured conceptually through a new social and political understanding of the manufacture and pricing of *options* and not simply posed as the conventional radical demand for publicly financed and administered social programs. »[1] This will be our starting point.

Adopting the language and conceptual framework of financial theory to address issues of compounding historical injustice starts with understanding finance as *a technology to manufacture liquidity*. Liquidity is a highly ambiguous and fleeting concept. As with many other things in finance, it is also highly self-referential. Liquidity corresponds to the ease with which an asset can be converted into money at a given market price. Or in other words, it describes the degree to which an asset or security can be quickly bought or sold in the market without affecting the asset's price. If something can be predictably sold at a certain price without having excessive impact on the price of other related assets, then it means that there is liquidity.[2]

Liquidity, then, can be understood as a *promise of reversibility*. The investor brings her money to the market, under the condition that she can take it back whenever she wants. Liquidity is the other face of trust, or rather the peculiar type of trust that must reign in financial markets for them to be effective. Confidence makes the market liquid and liquidity makes the market confident. Or in the words of Jim O'Neill, from Goldman Sachs: "Liquidity is there until it is not - that is the reality of modern markets."[3] All the financial innovations we've seen in recent years aims at increasing the liquidity of the market, transforming private debtor-creditor relation into something that exists publicly and can be bought by a third party. This operation is called *securitization*, and by some historical irony that should not evade our scrutiny, it is closely associated with a steep augmentation of budgets for private security

[1] Robert Meister, *Justice Is an Option: A Democratic Theory of Finance for the 21st Century*, University of Chicago Press, Chicago, 2020, p. xii (slightly modified version based on a previous version of the manuscript).

[2] "Market liquidity refers to the extent to which a market, such as a country's stock market or a city's real estate market, allows assets to be bought and sold at stable prices. Cash is considered the most liquid asset, while real estate, fine art and collectibles are all relatively illiquid."
https://www.investopedia.com/terms/l/liquidity.asp

[3] Quoted in Massimo Amato and Luca Fantacci, *The End of Finance*, Polity Press, Cambridge, 2012, p.16

services. This fact highlights something about the (anti)social nature of liquidity that John Maynard Keynes noted already in 1936:

> "**Of the maxims of orthodox finance none, surely, is more anti-social than the fetish of liquidity**, the doctrine that it is a positive virtue on the part of investment institutions to concentrate their resources upon the holding of "liquid" securities. It forgets that **there is no such thing as liquidity of investment for the community as a whole.**"[4] (my emphasis)

Global wealth, that is, the cumulative value of the world's assets, cannot be accumulated in financial form without also remaining liquid. That's the *raison d'être* of financial options and derivatives. Wealth must keep on moving. *Liquidity must go on.* This is the name of the financial game. In this sense, Meister suggests, we need to conceive of liquidity as the abstract form of absolute power of financial capitalism. This abstract relational imperative becomes particularly crucial when considering the massive bail outs put in place during the financial crisis of 2008, but also, more recently, following the beginning of the COVID pandemic in Spring 2020 (which was roughly 3-4 times bigger than the 2008 one) to restore trust in the credit market. Without these massive financial interventions, the financial markets would have crashed, precipitating a series of blockages or *illiquidity events* of almost inimaginable magnitude. In both cases, and Meister is adamant on this point, States have insured the liquidity of capital market *for free*, i.e., without requiring what he calls a *liquidity premium*. In an insurance contract, the risk is transferred from the insured to the insurer. For taking this risk, the insurer charges an amount called the premium. In the case of the bail out of 2008 and 2020, the governments have insured liquidity on the markets without

[4] John Maynard Keynes, The General Theory of Employment, Interest and Money, https://www.marxists.org/reference/subject/economics/keynes/general-theory/ch12.htm
The passage continues: "Conversely, from the standpoint of the financial community and hence of investors en bloc, - that is, from the standpoint of their consolidated balance sheet – investments are not liquid, since that would presuppose the liquidation of all economic assets in which financial capital is invested."

any premium, that is, without making any specific claim on the upside of the recovery. Meister's analysis leads to an unmistakable conclusion: « I believe that the production of liquidity should be the focus of socialism — that liquidity is not free, that it's not a positive externality, that it is something for which a political price can be extracted.»[5] In other words, preserving accumulated wealth in the form of financial assets has a price. And technically speaking, this price is that of the liquidity premium that should be requested by the States when they bail out the capital markets. This is the price of *justice as a (financial) option.*[6]

Meister is keen to point out how political these financial interventions are in effect. Quoting Roosevelt's brain-truster Adolf Berle, he writes:

> «'Any investigation of liquidity is a study of the mechanisms which make particular forms of wealth *acceptable*.' This implies, as Berle goes on to explain, that the political unacceptability of a particular form of wealth would preclude state support for its liquidity, and thus its convertibility into money and all the things that money can buy.»[7]

[5] https://democracyparadox.com/2021/10/12/robert-meister-believes-justice-is-an-option/

[6] The actual sum of the liquidity premium, according to Meister's own estimates, would have been, for the bail out of 2008, just short of the equivalent of that year's GDP, i.e. around 9 to 13 billions. In proportion, it is generally understood that the 2008 bail out was worth around 1 trillion for the United States alone (experts don't agree on what exactly should be included or not in this astronomical sum). Estimates for the 2020 COVID induced total stimulus package for the US varies from 3 to 4 trillions, while the whole world combines for an amount north of 10 trillions (and counting), according to the infamous consultant firm McKinsey:
https://www.mckinsey.com/featured-insights/coronavirus-leading-through-the-crisis/charting-the-path-to-the-next-normal/total-stimulus-for-the-covid-19-crisis-already-triple-that-for-the-entire-2008-09-recession.
These numbers need to be taken with a grain of salt - I'm only providing them to give a sense of the magnitude of these unprecedented financial interventions. To be clear: the issue here isn't about the amounts themselves, but about how and who they benefit in the first place. For an informative overview on the world's global wealth and the what it is made of, I can only recommend this most pedagogical diagram provided by Visual capitalist: https://www.visualcapitalist.com/all-of-the-worlds-money-and-markets-in-one-visualization-2020/

[7] Robert Meister, *Justice Is an Option: A Democratic Theory of Finance for the 21st Century*, p.139

The issue of how liquidity is guaranteed in current capital markets by democratic States is thus a highly political one, yet one that, for different reasons, we struggle to address fully as such. And more importantly, for the purpose of this volume, it also represents one key perspective on how to envisage the macro-financing of something like a UBI or some other justice-oriented wealth redistribution program.[8] But if the manufacturing of liquidity depends on how everyone of us is used as a collateral in complex financial arrangements and thus truly is, at the end of the day, a question of social and political acceptability, doesn't it make it even more attractive – and even more so, necessary – to simply *occupy everything*, that is, to put to a (definitive) halt this whole slick and abstract financial machinery and the logistical apparatuses it depends on, thus triggering a cascade of ever-amplifying illiquidity events? This question bring us back to another one, stemming from the incandescent core of the Occupy Wall Street and which remains largely unanswered to this day: *how to occupy a (financial) abstraction?*

Meister directly addresses the problem of what he frames as « the payoff of the revolutionary option » in these terms:

> «For the purpose of funding justice, the most difficult abstract question is what the payoff of the revolutionary option would be. The ideal of revolutionary abundance assumes the accumulated wealth would be preserved intact despite its redistribution. In contrast, the ideal of revolutionary asceticism assumes that it shouldn't matter if accumulated wealth would lose all its value by being redistributed, since this would only prove that it was never real. **The truth is that we don't know to what degree asset prices—essentially, the liquidity of capital markets—would recover under the, presumably, revolutionary state**

[8] «Put very crudely, it seems to me that the collective demand for money – neither wages nor credit, but simply money as a redistribution of wealth – could be disruptive of the financial system in the sense of **making a common claim on the publicly created and guaranteed collateral that is used to secure accumulated wealth that remains in private hands**.» (emphasis added). Robert Meister, "Liquidity", in Benjamin Lee and Randy Martin, *Derivatives and the Wealth of Societies*, Chicago, Chicago University Press, 2016, p.173.

> **the world** in which asset ownership and/or the flows of revenue and collateral were reallocated.»[9] (my emphasis)

Meister's core political concern articulates around how to insure the conditions of political effectuation of historical justice *in the now,* even in times in which the option of revolutionary illiquidity can't or should strategically not be exercised.[10] This is why he suggests to make use of the tools and language of finance, and especially the option pricing theory framework, *to characterize revolutionary justice as a financial option.* An option is, roughly put, a way of attaching a present value to something that is unknown in the future. In financial markets, options consist of derivative products that give buyers the right, *but not the obligation*, to buy or sell an underlying asset at an agreed-upon price and date.[11] As explained earlier, there is a price that is attached to the creation of financial options themselves - a premium.

Following this logic, and in elegant speculative symmetry with his understanding of finance as a technology to manufacture liquidity, Meister therefore presents democracy as a *technology to manufacture alternatives to revolution.* We know that, historically, the Welfare State emerged as an institutionalized class compromise following decades of hard-fought struggles, leading to what came to be known as a «social pact».[12] It goes without saying that the social progress resulting from this

[9] *Ibid.*, p.230.

[10] «I'm saying that historical justice can have value even when the option of revolutionary illiquidity can't be exercised, and that its value can be measured by the premium that could be charged for government-provided liquidity in capital markets — the "liquidity put."»
https://www.salon.com/2018/07/08/scholar-robert-meister-on-a-new-model-using-the-financial-markets-to-fuel-historical-justice/

[11] «Options are versatile financial products. These contracts involve a buyer and seller, where the buyer pays a premium for the rights granted by the contract. Call options allow the holder to buy the asset at a stated price within a specific time frame. Put options, on the other hand, allow the holder to sell the asset at a stated price within a specific time frame. Each call option has a bullish buyer and a bearish seller while put options have a bearish buyer and a bullish seller.»
https://www.investopedia.com/terms/o/option.asp

[12] https://www.jacobinmag.com/2021/05/welfare-state-class-struggle-confrontation-compromise-labor-union-movement

fairly long-lasting arrangement between labor and capital depended on a strong labor movement. And it's also quite obvious that after 40+ years of neoliberal offensive and ruthless financialization of our economies and social relations, the balance of power has shifted dramatically. Social struggles are crucially in need of updated political strategies. In the last instance, Meister's goal is nothing less than to offer « a discursive framework and political practice for the justice-seeking subject in the age of financialization, in the way that Marx did for the justice-seeking subject during the industrialization of manufacture.»[13]

Concretely speaking, If the Welfare State was the political price exacted for *not* exercising the option of a General Strike, argues Meister, there should also be a political price for *not* exercising the option to bring on a liquidity crisis, let's say, through the means of a collective action aimed to occupy (i.e. re-possess) collateral that the financial system is laying claims on. In other words: «Critically appropriating the language of financialization thus allows us to see more clearly how democracy can reintroduce the political risk that government will not restore liquidity to capital markets when they need it most.»[14]

But what are our options exactly? What does it mean for a justice-seeking subject in the age of finance to occupy *this* time and space, to collectively incorporate the strike price of justice, knowing how notoriously difficult it is to challenge the dominion of financial abstractions governing our lives at a distance? Or again: if we are to bracket, for the time being, the actual recourse to direct action generating illiquidity events, then *what kind of otherwise liquidity-making practices can we imagine that wouldn't end up reinforcing the prevailing power relations?*

Meister's interpretation of option pricing theory generates a pedagogy of (il)liquidity that plunges its theoretical roots in a reading of Walter Benjamin's *Thesis on the Concept of History*. It aims at reanimating the revolutionary urgency at the core of Benjamin's political project with a surprising yet highly rigorous financial twist that emphasizes the now-time of justice:

[13] From private conversation.

[14] *Justice Is an Option: A Democratic Theory of Finance for the 21st Century*, p.11

> A proper description of our task is to understand why historical injustice is rarely redeemed, and yet must remain redeemable, and then to describe, as Benjamin tried to do, the exceptional (miraculous) status of a "now-time" in which another time is also made present and thus redeemed. This intertemporality of justice is not merely a matter of occluding the history's apparent losers but of the proper valuation of the present claims that can be made through them – and of **seizing a moment when that value is finite and calculable.**»[15] (my emphasis)

This way of dramatizing *the financial time that remains* is highly speculative. All the more so when it comes to the question of *intertemporality of justice*, a key component of Meister's thesis that I won't be able to fully address in the context of this article, even though it directly concerns the very existence of the many worlds in their financial plurality.[16] What I find most stimulating in Meister's theological-political insistance on the finite and calculable is that instead of cloaking itself into the moral mantle of political infinitism, the idea of justice as an option

[15] Robert Meister, *After Evil. A politics of Human Rights*, Columbia University Press, New York, 2012, P.248. "The real challenge is to develop a financial model that explains how the constructive value of unjust enrichment fluctuates over time as the political, social, and economic relations of the affected groups also change. This poses Benjamin's question of when to seize the present moment to redeem the past. Here, however, **redemption would, arguably, take the financial form of a preference on the part of both victims and beneficiaries for liquidity rather than running debt.** » (p.247) (my emphasis)

[16] "Optionality of the kind that finance illustrates is more broadly about synchronizing heterogeneous temporalities, indexing heterogeneous cultural discourses, tokenizing the relative rates of change within and among heterogenous systems of valuing and ranking—the list could go on. Such forms of heterogeneity no longer need to be reduced to a General Equivalent if liquidity can be added through options that can index their changes to those in other, disparate, value realms." Robert Meister, *Justice Is an Option. A Democratic Theory of Finance for the 21st century*, University of Chicago Press, Chicago, 2021, p. XXX. When articulating this idea, Meister references this article from Economic Space Agency, "On Intensive Self-Issuance: Economic Space Agency and the Space Platform," in *Moneylab Reader #2: Overcoming the Hype*, ed. Inte Gloerich, Geert Lovink, and Patricia de Vries, Institute of Network Cultures, Amsterdam, 2018, 232–42. http://networkcultures.org/wp-content/uploads/2018/01/21-ecsa.pdf

fully assumes the core enabling constraint of finance, namely, the fact that finance is, effectively and etymologically, about how we deal with « endings », about how we make ends meet.[17] Finance is indeed a system that constantly presupposes its own catastrophic end, and benefits from how long it can be delayed. To be sure, many worlds have come to an end before ours – just ask indigenous people all around the planet. Why is it so hard to fully acknowledge the constitutive limits of the world(s) we live in? Finance is traditionally understood as the transformation of radical uncertainty into manageable risk. But as it prospects for ways of generating surplus-value, the virtual body of capital generates highly qualified relations to futurity that challenges the very limits of our cultures of knowing and forecasting. This speculative movement calls for new ecologies of practices and knowledges to account for our current economic abstractions. How can we leverage our own capacities to take risks and enter into metastable collective compositions beyond what is deemed possible – or insurable?

The passage from a logic of risk management to worlding practices of shared metastability is, I would argue, a core component to reclaim and, eventually, decolonize finance as we know it. Liquidity irresistibly flows toward the one world of Capital. Or rather, it is its most concrete, yet inherently abstract, manifestation. Inversely, we need to imagine a *cosmo-financial pluralism* that doesn't simply take for granted the alleged superiority of the « commons » as a generic ethical, political and organizational horizon, but engages in inventing transversal manners of accounting otherwise that do not shy away from addressing the difficult question of the *(in)commensurability of value claims* through and between the many worlds.

[17] Peter Sloterdijk provocatively describes political infinitism as follow: « Political infinitism, which is the political definition of the left, has so far had to distance itself from all the rhetoric and practice of concrete community, because it requires a politics of the finite. Alain Badiou has recently reformulated the axiom of a postmarxist politics of emancipation: "the situations of politics are infinite". False but clear: by reading it, one understands well that the metaphysical left proposes the infinite as a critique of the finite — which reveals the religious roots of any left politics of the possible and the real. (...) On the other hand, the piquancy of recent communitarianism is to clarify the conditions of a left politics of the finite.» *Sphères II: Globes. Macrosphérologie,* Paris, Librairie Arthème Fayard, 2010, p. 362. (my translation)

Cosmopolitics of the kind developed by thinkers like Isabelle Stengers, Bruno Latour or Felix Guattari is concerned with more-than-human communities and the way they attune with their associated milieus. The cosmo-financial proposal extends this view by integrating the promises and challenges raised by cryptoeconomics and derivative finance's affordances for new collective incorporations of value. *The cosmo-financial art of belonging in becoming foregrounds value discovery processes that are not confined to the logic of the market*. For what we owe to one another is not something in particular: it is the very unknown that envelops our existences, the zones of opacity and indetermination delineated by our more or less felicitous encounters. The *cosmo*- in cosmo- politics/technics/finance refers to the unknown constituted by these multiple, divergent worlds and to the articulations of which they are capable of.[18] At this level of analysis, we start seeing better how any UBI system always also imply an ethico-aesthetical appeal to something like an IBU, that is, *Intensive Basic Units* made of wildly (or magically) interested dividuals actively curating stakes into one another (ad)ventures, facilitating in all fashions the haptic experience of feeling through others (IBU is always already an "I Be You" :).

Theoretically speaking, the perspective of the cosmo-financial relies on a concept of value that foregrounds *the active energy of discrete assemblages*. This focus on the collective incorporation or embodiment of value is decidedly future-oriented. It is speculative and pragmatist in scope, as it diagonally cuts through, if only for polemic purposes, the perennial debate around the dialectic of exchange VS use value (the inescapable trope of many a zombie marxist), foregrounding instead a different set of concerns around the financial art of asset formation and new types of equity-based relations. This emphasis on emergent worlding practices with a financial edge also tends to displace the emphasis put on the question of debt, a perspective largely hegemonic in critical academic and activist circles. As suggested by Meister and others, finance actually offers many others tools of analysis to challenge and reconfigure

[18] For more on this question, see Erik Bordeleau, « After the Attention Economy: Notes Toward a Cosmo-Financial New Serenity », *in 2038 – The New Serenity,* German Pavillion of the 17th International Architecture Exhibition, Venice Biennale*,* Sorry Press, Munich, 2021*. https://www.sorry-press.com/2038-the-new-serenity*

our subjection to capital. In this sense, if UBI is indeed a cosmogenetic technique as suggested in the UBI Manifesto redacted by the Institute for Radical Imagination, I believe we need to explore in what way the prefix *cosmo-* calls for worlding practices that could more openly assume the activating powers enclosed in the financial art of asset formation.

Long before Deleuze wrote about the question of belief in the world in cinema, he was already prefiguring a politics of choked passages and (un)timely contractions, conceiving of artful differences as introducing a «freedom for the end of *a* world».[19] In the same spirit, I would say: *other financial ends of the world are possible*. But in order to decolonize finance from within, and especially so in the context of the affluent West, we need to realistically – if only speculatively – start to flesh out the type of claims to abundance we imagine for ourselves and the generations to come. And we need to do so not only in the mode of an infinite demand for income redistribution addressed to the State, but in a way that integrates the resources of financial theory to inform otherwise worlding practices.

In guise of conclusion, I would like to rapidly indicate two prospective contributions – one fictional, one IRL – pointing in that direction. In his suggestively entitled *Another Now: dispatches from an alternative present* (Penguin Book, 2020), ex- Greek minister of finance Yannis Varoufakis presents a thought-provoking speculative financial fiction that significantly contribute to opening up the gate of radical imagination for otherwise financial worlding practices. Varoufakis's book reads as what a fellow sci-fi writer with financial inclinations, Kim Stanley Robinson, describes as *optopia*, that is, something like « the best scenario one can still believe in ». Varoufakis's Other Now happens in a slightly bifurcated universe accessible from our current present through some special warpzone technology. It unfolds as a speculative thought experiment in which a series of pragmatic and visionary measures have been implemented to redress historical injustices. In the Other Now, for instance, tradable, corporate shares have been abolished, « thus damming

[19] Deleuze, Gilles (1994) *Difference and Repetition*, trans. Paul Patton, Columbia University Press, New York, p.293. For a full analysis of this passage on which concludes *Difference and Repetition*, see my «A Redemptive Deleuze? Choked Passages or the Politics of Contraction», *Deleuze Studies Journal*, 8:4, Edimburg University Press, 2014.

the whirlpool of financial speculation until its torrent is reduced to a tepid stream»[20]; yet, a variety of equity-based mechanisms and stakeholding relations proliferate, alongside a multitude of community currencies, suggesting a whole array of alternative liquidity-making practices. Indeed, one of the most interesting aspects of Varoufakis' book is how seriously (and practically) he engages in describing the inner functioning of this otherwise economy he sometimes describes as *market without capitalism*.

The Other Now emerges out of a series of political interventions that have generated illiquidity events of massive scale. One of them articulates around what Occupy Wall Street has or could have been. The problem with Occupy Wall street, writes Varoufakis from the perspective of the Other Now, is that occupying spaces to reclaim capitalism is futile, since «capitalism doesn't live in space but in the ebb and flow of financial transactions.»[21] In the Other Now, this diagnosis translates into a movement of financial activists led by the *Crowdshorters*. By successfully convincing a critical mass of people to all default on their utility bills at the same time, they were « the first group to demonstrate the vulnerability of financialized capitalism and the power of a well targeted digital rebellion. »[22] It is interesting to note that this fictitious yet highly plausible scenario corresponds almost integrally to what the great political economist Dick Bryan has presented in different occasions as a proposition for a *Household Union*. For Dick Bryan, just like for Robert Meister when he insists on financial theory's specific affordances for determining finite yet fluctuating values in time, it is essential that «we talk about what is happening to households in new ways. *It's not an income distribution issue; it's a risk distribution issue*. We usually don't have the language (at the level of households) to talk about that. Financeers do. So they are able to impose risk because they understand the process. We are bunnies in the headlights of risk transfer.»[23]

[20] Yannis Varoufakis, *Another Now: dispatches from an alternative present*, Penguin Book, London, 2020, chapter 4, "How Capitalism Died".

[21] *Another Now: dispatches from an alternative present*, "chap.4, "How Capitalism Died".

[22] *Ibid.*

[23] https://www.youtube.com/watch?v=bEj89w-mZNQ

Early on in the alter-financial novel, one of the protagonists asks his double in the parallel world a key question: how capital is formed and accumulated without a stock market? The Other Now is described by the alter ego as a regime of «democratized inequality» in which every citizen is provided at birth with a bank account called «Personal Capital» (or *PerCap*) that includes three strictly separated types of fund: Accumulation, Legacy and Dividend. Salary and bonuses related to work are credited in the Accumulation fund, roughly the same way as it is today. Legacy is more innovative: everyone at birth is credited with an amount of money that can only be invested for productive activity purposes. «Babies are still born naked but every one of them comes into the world with a bundle of capital provided by society. This means that when they come of age and are ready to enter an existing business, or start one alone or with others, every youngster has some capital to deploy.»[24] The third fund, Dividend, is presented as an upgraded version of a universal basic income. Every month, each citizen receives a monthly payment that «liberates everyone from the fear of destitution», providing «people who do not care to engage in business activity with sufficient income to provide priceless contributions to society»[25], something that the protagonist provocatively describes as «a right to laziness». Varoufakis insists in contrasting Dividend from more traditional UBI scenarios: «The key was that Dividend was not financed by taxation; it was, rather, a real dividend that people received as co-owners of the capital stock they were collectively producing – even if they did not do what we readily recognize as work.»[26] *Another Now* is keen in suggesting how active stakeholding in re/productive activities can be envisaged as a viable alternative to the corporate world structured around shareholding structures that are by essence ecologically unsustainable and extractive, allowing the reader to imagine a series of mutually reinforcing circles of activating reciprocity. Or as we playfully like to say in The Sphere, a research-creation project

[24] *Another Now: dispatches from an alternative present*, "chap.4, "How Capitalism Died".

[25] *Another Now: dispatches from an alternative present*, "chap.3, "Corpo-Syndicalism".

[26] Chap. 4, « How Capitalism Died ».

exploring new ecologies of funding for the performing arts: *you can't be alone in a liquidity pool!*[27]

The possibility offered by blockchain technologies to participate in the design of new protocols for networked asset formation points towards way of renewing our collective incorporations of shared lived abstractions, i.e. the way we come together without becoming one, generating derivative value along the way. This concrete utopia is alive and well in the world of web 3.0: a myriad of monetary self-issuances that could be modulated at will, following the affordances of a given ecosystem and in response to the inter-species web of entanglements in which they are embedded. For, to paraphrase Donna Haraway's provocative and staying-with-the-trouble insight: it matters what worlds world worlds; and *it matters what measures measure measures*.

Circles, the Berlin-based project for a trans-local UBI network, is a good example of such proliferation of new forms of plural organizing experimenting with money as a medium for collective incorporation.[28] Circles is an original initiative leading the way toward what they call a *Money commons, i.e.* a confederation of local community currencies aiming to operate a civilizational paradigm shift in how we resist monetary extractivism by keeping the value produced locally within the community. What is particularly interesting about *Circles* is that it doesn't presuppose what or *who* a community is from the outset. Rather, it constitute itself as a collective power to redesign the economic relations we are embedded in – a form of curated yet inclusive network based on an expanding web of trust. Contrary to most UBI propositions, *Circles* doesn't address itself to the State as purveyor of income of last resort (although it still needs some massive support to bootstrap initial liquidity – in this case, the funding is provided by a generous sponsor coming from the blockchain world). Circles empowers its participants to design a pluriverse of claims on the already existing wealth in society, reintegrating it into circuits of mutually addressed promises. Inspired by the work of David Graeber among others, Circles exemplifies how, at

[27] https://www.thesphere.as/ This was actually the name of the Sphere' Cryptoeconomic Design Lab held on April 22-23rd 2021 (check the Sphere Timeline on the website for more details about the event).

[28] https://handbook.joincircles.net/docs/users/

the end of the speculative day, money is nothing else than an IOU, an «I Owe You», a document that acknowledge the existence of a debt. The value system generated by *Circles* reflects this state of fact in its design by allowing people to issue promises unconditionally, and decide in which sort of relationships they want to be involved in and how.

Can we imagine the scaling of such a world in which a myriad of quality-charged currencies meet with one another, each of them carrying the senses and flavours of the community issuing and backing them? These different self-issued tokens would be a bearer not only of monetized value, but also an index of local expressive forces. These new modes of measuring collective outputs would catalyse new calibrations between the realm of the quantitative and the realm of the qualitative, providing a unique answer to the proverbial interrogation about what money can and cannot buy.

Finance as an expressive medium commands a logic of implication. Self-issuance is about exposure to an outside, but it doesn't necessarily mean a full-fledged exposure to the full contingent outside of the market. The advent of blockchain and distributed ledger technologies is but one new chapter in a long and complex history of record keeping, archiving practices and institutionalized trust that goes back to the origin of writing itself. One thing is for sure: whatever techniques we use to keep ourselves accountable, something always exceeds. Anarchic shares will proliferate away from the grid. You can get a hold on it as long as you pass it on. We live beyond our means and our ends, we set them free so they take us with them and this fills us with a strange joy, for we owe each other the indeterminate.

Political Correctness: the Twofold Protection of Liberalism

Sandra Dzenis and Filipe Nobre Faria

Dzenis, S. & Nobre Faria, F. (2020). Political correctness: the twofold protection of liberalism.
***Philosophia 48*, 95–114.**
https://doi.org/10.1007/s11406-019-00094-4

1. Introduction

As a term, political correctness (PC) is regularly present in the political lexicon of the contemporary West. The term now refers to a concrete social phenomenon with broad recognition. Although defining PC can be contentious, according to the Oxford Dictionary of New Words, PC is ‘conformity to a body of liberal or radical opinion on social matters, characterised by the advocacy of approved views and the rejection of language and behaviour considered discriminatory or offensive’ (Knowles and Elliott 1997). As we understand it, this definition adequately describes the nature of the phenomenon; hence, it is the reference throughout the article.

Even though it is possible to trace back the term’s origin to historical forces such as Marxism and Maoism (D’Souza 1991; Hildebrandt 2005), many critics of the term understand its modern meaning as an invention of the political Right to marginalise the Left’s efforts to reach a more egalitarian society (Feldstein 1997; Sparrow 2002; Wilson 1995). These critics think the political Right invented the concept of PC to strengthen the right to dominate women and minorities, including racial minorities and homosexuals. Whatever its origins, the term is more popular among detractors of the content of PC. Accusing someone of PC ‘has become a sarcastic jibe used by those, conservatives and classical liberals alike, to describe what they see as a growing intolerance’ that shuts down debate with accusations of ‘sexism, racism and homophobia’ (R. Bernstein 1990).

While some political egalitarians defend certain aspects of PC (Fish 1994), others accept parts of the conservative critique (Gitlin 1995; Lea 2009). Some others, like Richard Rorty (1998, pp. 81–82), regard PC as a product of civilization that reflects ‘a basic desire to tolerate, not persecute, those who have different faiths, beliefs, or skin colour’ (Roxburgh 2002, p. 302). Although perhaps few people would claim to be believers in PC as a label, the term widely represents the advocacy of censorship that aims at protecting vulnerable groups. In social discourse, the term identifies a practice with ideological advocates regardless of whether they identify with the PC label. In this sense, the social impact of the term justifies the choice of PC over less impactful terms such as self-censorship (Cook and Heilmann 2012) or conformity (Williams 2016).

Debates about PC have focused on the consequences PC brings to academic freedom and political discourse (Bloom 1987; Cole 2006; Kimball 1990; Lukianoff and Haidt 2015; Moller 2016; Williams 2016). Within this overarching socio-academic debate, we detected a general argumentative trend that divides claims in favour and against PC into two epistemic and normative camps characterised as follows:

(1) the post-modern-like advocates of PC who favour regulating speech and behaviour to achieve social justice (Fish 1994; Lawrence 1990; Matsuda 1993) and
(2) the Enlightenment liberals who oppose PC by upholding truth-seeking open discourse and scientific rationality (Chait 2015; Furedi 2016; Pinker 2017; Pinker 2018; Rauch 2013).

This dichotomy arose when the second camp denounced the first one and adopted a comprehensive anti-PC stance.[1] The aim of this paper is to show that this dichotomy does not hold up under scrutiny. We argue that

(1) no visible camp is in fact defending an open-ended scientific discourse, that (2) PC is a protection mechanism of liberal values and that (3) both sides represent PC.

The current debate is in reality about how to protect liberal values.

The structure of the argument is the following. First, the paper traces back the ideological roots of PC to core liberal values and goes on exploring how PC's function is to protect and further liberalism. Then,

[1] The two camps battle over epistemological differences. If there is no substantive truth – as in post-modernist discourse –, potential claims for tolerance of offensive viewpoints in the name of truth and open discourse may lose their value. In this sense, the normative clash between the two camps involves asserting the epistemic status of scientific truth. Not everyone must identify with one of the two camps. Some may, for instance, support certain levels of PC while upholding scientific realism. Yet, supporting PC within science requires science-based supporters of PC to justify why truth is less relevant than restricting viewpoints for moral reasons, which again brings the epistemic dimension to the fore. Epistemology is key in the current PC debate.

it claims the post-modern abandonment of enlightened truth-seeking is a particular form of PC, which attempts to protect liberalism from illiberal forces. Afterwards, the paper argues that although Enlightenment liberals claim to oppose PC, they still impose it by only engaging with truth-claims within the liberal framework. At the close, we show that science does not commit itself to liberalism.

2. PC as a Mechanism to Protect Liberty and Equality

PC emphasises a strong inclusive position, according to which individuals require moral equality in all aspects of life regardless of their religion, race, age, ethnicity, sex or gender. To enforce this attitude, PC advocates may use affirmative action or restrict free speech with speech codes and anti-discrimination laws (Bernstein 2003, pp. 1–4). However, Glenn Loury suggests that PC also implies conformity to a desired opinion on socio-political matters, which proliferates via social pressure:

> (…) the more subtle threat is the voluntary limitation of speech that a climate of social conformity encourages. It is not the iron fist of repression, but the velvet glove of seduction that is the real problem. Accordingly, (…) the PC phenomenon [can be treated] as an implicit social convention of restraint on public expression, operating within a given community. (…) Members whose beliefs are sound but who nevertheless differ from some aspect of communal wisdom are compelled by a fear of ostracism to avoid the candid expression of their opinions (Loury 1994, p. 430).

Given the existing social pressure for conformity of beliefs, a scrutiny of the ideological or moral underpinnings of PC is of importance to understand its ontological constitution. In particular, there is a strong connection between politically correct (pc) attitudes and specific liberal values, such as individual freedom and equality, to an extent that it is possible to understand PC as being underpinned by liberalism.

As Michael Freeden points out, although the fluidity of liberalism may force us to acknowledge its existence in the plural, 'liberalism is a particular configuration of political concepts that has a loose but identifiable morphology' (Freeden 2008, p. 12), of which liberty and equality are identifiable central values common to all liberal versions. This specific liberal morphology makes it possible to address liberalism in the singular, thus distinguishing it from other value systems. Still, PC seems to fit better with a vertical conception of liberalism – also understood as welfare or social liberalism – which promotes upward social movement and relies on positive conceptions of freedom. Positive freedom emphasises the need to remove the inhibitions of any social structure that prevent individuals from exercising their free will, inhibitions such as economic hardship, classism or racism. Such a conception of liberalism understands that unhindered self-realisation is illusory and therefore achieving real freedom and equality requires communal and state assistance. The pc approach also relies upon the idea that individuals of marginalised groups require assistance from community and state when struggling against offensive attitudes. Censoring offensive speech and attitudes that hinder the freedom of these individuals can be a way to free them from oppressive social structures.

A horizontal conception of liberalism – also regarded as constitutional or classical liberalism – seems to be less amenable to PC. This conception of liberalism emphasises free choice, dispersed knowledge and constitutional protection of negative liberties. Negative liberty is freedom from external restraints on the actions of individuals, something associated with minimal state representations. According to these representations, authority focuses on protecting direct harm and not on removing structural obstacles to achievement. The communal and state actions that would legitimise PC under a vertical/social conception of liberalism sit uneasily upon the horizontal/classical conception. The reason being that the latter conception relies on negative freedom. Thus, the laissez-faire attitude coming from a negative conception of freedom is more amicable to uncensored speech. Yet, as we will show later in the article, even a more horizontal/classical liberal position can use PC to defend liberalism. It can do so by endorsing a culture of voluntary ostracism towards illiberal viewpoints.

Subject to the condition that – overall – legal equality is in place in liberal democracies, pc attitudes focus mostly on substantive or enabling equality, an equality that aims at levelling departure points and enabling achievements. A pc position also emphasises the freedom of the individual, meaning every individual should be free from constraint to pursue one's own notion of a 'good life' (Rawls 1993, p. 19). This ethical pluralism entails that different notions of the good life are of equal value as long as these notions respect basic universal freedoms. And it ties to liberal pluralism in the sense that free individuals should be eligible to follow their own perception of the ethical life.

In consequence, a pc attitude manifests itself by assuming that the desired freedom for individuals in society is attainable by implementing not only formal but enabling equality (e.g. affirmative action, women's quota, etc.). In addition, pc positions inhibit the accentuation of certain individual and group differences to prevent unequal treatment. For instance, it is pc to deny or at least downplay innate human differences because these differences may explain inequalities of outcome (e.g. the gender pay gap). Pc thought seems to rely on the assumption that the way to achieve the most significant goal of individual freedom is through (a certain type) of equality.

Scholarly literature suggests that liberal thought has its foundation in the legacy of the Enlightenment (Brink 2000; Byrne 1997; Waldron 1993; Zafirovski 2011). As noted by Bert van der Brink (2000, p. 13), the idea of equality rests upon the liberal notion that all human beings hold the fundamental right to respect due to their status as reasonable and free individuals. The belief that the individual mind can gain genuine knowledge and grasp the fundamental principles of the world led to the conclusion that we should treat all reasonable beings as equals. As John Locke argued, nobody should ever be 'subjected to the Political Power of another without his own Consent' (Locke 1988, II, sec. 95) given men's moral sameness in nature.

Likewise, the Enlightenment gave birth to human rights to protect the autonomy and equal liberties of individuals. Perhaps the practical implementation of some human rights requires a certain level of PC. How, for instance, can one expect ethnic minorities/LGBT members/disabled people to take part in the cultural and political life of the community (UDHR, 2010, art. 27(1)) if they feel marginalised by some members of society? Thus, PC advocates campaign for speech codes and for conformity of thought towards minority groups in order for these groups to enjoy their complete human rights. Also, the right to education (UDHR, 2010, art. 26(1)) may lead PC proponents to the conclusion that only with the help of affirmative action can certain minority groups enjoy their rights. Even trickier seems to be the right to liberty (UDHR, 2010, art. 3). Some PC supporters claim unrestricted speech and discriminatory behaviour threatens the liberty (and therefore a major human right) of affected human beings (Delgado 1982; Matsuda 1989; Parekh 2017).

3. PC as Liberating Tolerance

The most common advocacy of PC comes from contemporary post-modernists and critical theorists (Fish 1994; Lawrence 1990; Matsuda 1993; Rorty 1998) who often advocate forms of post-modern liberalism (Dryzek 2000, p. 27). But what is the standard intellectual source of this advocacy within PC-focused scholarship? When starting his essay 'Imagined tyranny'? Political correctness reconsidered, sociologist Paul Hollander puts forward the concept of repressive tolerance – introduced by Herbert Marcuse (1965) – which is of high influence for 'the most widespread form of institutionalized intolerance in American higher education' (Hollander 1994, p. 51). Also, in their work The shadow university: The betrayal of liberty on American campuses, Kors and Silverglate (1998) argue that Marcuse's philosophy is the intellectual progenitor of PC at university campuses: 'The contemporary movement that seeks to restrict liberty on campus arose specifically in the provocative work of the late Marxist political and social philosopher Herbert Marcuse', who challenged 'the essence and legitimacy of free speech' (Kors and Silverglate 1998, p. 68). It is thus significant to set out Marcuse's theory of repressive tolerance, which this scholarly literature shows to be the birth hour of PC.

In his essay on repressive tolerance, Marcuse tries to figure out if there are ethical limits to tolerance and what consequences come from this enquiry. According to Marcuse, universal tolerance is only real when serving the cause of liberation and proper tolerance cannot arise as long as the holders of power and the guardians of the status quo indoctrinate society to keep inequalities stable. He considers it unfair to let the powerful and the powerless play under the same rules, because the powerful would always win and, as a result, would impose a violent and repressive agenda on the powerless. Hence, he points out that movements from the Left must replace the political Right. This replacement aims at implementing the Left's 'liberating tolerance' (Marcuse 1965, p. 109), which censors oppressive speech while expelling the Right's repressive tolerance, namely the repression operating under the guise of free speech.

Marcuse asserts that liberating tolerance is the only way to exercise (civil) rights and liberties for the oppressed. Hence, it should 'be enforced by the students and teachers themselves, and thus be self-imposed', withdrawing any 'tolerance toward regressive and repressive opinions and movements' (Marcuse 1965, p. 101). As a result, Marcuse's liberating tolerance, under which real freedom could flourish, should thrive first on university campuses before the concept encroaches upon the greater society: 'This re-education alone could create a progressive society, where true freedom and democracy would reign' (Kors and Silverglate 1998, p. 71). While people outside academia may know little about Marcuse's formula for a progressive society, his prescriptions represent the paradigm for speech restrictions in the contemporary academic world. The liberal dimension of Marcuse‘s rhetoric is not always straightforward, perhaps because of his Marxist background. Yet, his philosophy suggests that universal liberties can only flourish within society if pc measures minimise the power and influence of any repressive establishment. Today's advocates of repressive tolerance are more explicit regarding the liberal aims of PC (Kernohan 1998; Levin 2010).

Contemporary social scientists advocating PC, such as Charles R. Lawrence, Richard Delgado and Mari Matsuda, build their research on race and gender bias upon Marcuse's idea of repressive tolerance (Delgado 1982; Lawrence 1990; Matsuda 1993); that is, on the idea that pc speech restriction applied to dominant/privileged groups allows for all members of society to experience equal freedom. Lawrence, for instance, notes that because white supremacy is the underlying message of racist speech, nonwhites experience limited life opportunities: 'There can be no true free speech where there are still masters and slaves'

(Lawrence 1990, p. 481). Matsuda adds that official tolerance of racist speech on campus is harmful since it attacks 'the goals of inclusion, education, development of knowledge, and ethics that universities exist and stand for' (Matsuda 1989, p. 2371). In addition, Matsuda argues that individuals do not depart from an equal point. As a result, evaluating hateful speech regarding race/ethnicity must take the targets of such speech into consideration. Delgado concludes that racial speech cannot be part of the marketplace of ideas because instead of informing or convincing the listener, racial speech merely inflicts harm. Hence, such speech prevents the speaker and the listener from having a meaningful discourse (Delgado 1982, p. 177). By denying unrestricted freedom of expression, Delgado desires effective freedom in order for 'all citizens to lead their lives free from attacks on their dignity and psychological integrity' (Delgado 1982, p. 181).

The arguments against robust free speech put forward by Lawrence, Matsuda and Delgado echo Marcuse's concept of repressive tolerance. In relation to implementing 'repressive tolerance' on campuses, Kors and Silverglate (1998) argue that university speech codes reflect Marcuse's idea of freedom and tolerance. They claim these Marcusian values try to balance the right of free speech with the right of not being harassed, to balance negative freedom with positive freedom. In this sense, speech restrictions assure liberty for some by limiting it for others.

Philosophers Andrew Kernohan and Abigail Levin, for instance, worry about state neutrality, which in the PC debate means unrestricted freedom of expression and a handsoff approach regarding the cultural marketplace. They argue that contemporary liberalism has given too much emphasis on tolerance at the cost of equality. Hence, there is a need for an advocacy strategy toward cultural reform, a compromise between unrestricted freedom of expression and coercive censorship by the state (Kernohan 1998; Levin 2010). In the same wavelength, Kernohan suggests state-promoted social conformity (i.e. PC). He points out that tolerance is not something for the enemies of liberalism to enjoy:

> Liberalism requires tolerance of all manner of views on how to lead a worthwhile life, but not of views that deny the fundamental assumption of moral equality. (...) Liberal

> tolerance comes to an end for views (that are) inconsistent with liberal principles, and [that] threaten significant harm to society as a whole. (…) Therefore the liberal state must take an active role in reforming culture and combatting the cultural oppression of groups (Kernohan 1998, pp. 4-25).

Overall, contemporary liberal academics, such as Lawrence, Delgado, Matsuda, Levin and Kernohan, support certain pc measures on behalf of the liberal state to counteract oppression and social inequalities. The specific claims of Kernohan and Levin suggest that PC operates as a mechanism to promote and defend liberalism.

Some may argue that because critical or post-modern PC defends rights on the basis of group identity, it deviates from liberalism's commitment to ontological individualism, therefore becoming illiberal. This claim grows stronger because some early proponents of PC, such as Marcuse, came from a Marxist-influenced intellectual sphere. Yet, PC is not an illiberal phenomenon by necessity. In fact, group identity is often a liberating argumentative tool that marginalised individuals use against any oppressive institution which discriminates against them because of their group identity. In this sense, in order for individuals of unprivileged groups to enjoy liberty and equality, they need to emphasise their identity as the reason for their lack of equal liberty. We are not dealing with a novel issue. Throughout history liberals have used collective-based and identity-based concepts, such as the people, to overthrow non-liberal and allegedly oppressive political regimes (Eddy 2017). Due to their flexibility, liberal values often accommodate their egalitarian critics. In the words of John Dryzek:

> Liberalism is the most effective vacuum cleaner in the history of political thought, capable of sucking up all the doctrines that appear to challenge it, be they critical theory, environmentalism, feminism, or socialism (Dryzek 2000, p. 27).

In particular, because egalitarian doctrines are many times in line with the moral desirability of liberal values, these doctrines can flourish within the fluid realm of liberalism. As for PC, the current and most common justification for its legitimacy relies on liberal concepts. Namely, speech restrictions are legitimate because they increase the liberties of individuals in marginalised groups by enhancing positive freedom, while these liberties deteriorate through negative freedom and unfettered critical discourse.

Not all authors following Marcuse's repressive tolerance may identify as liberals. Some would balk at applying the term liberal to their lines of thought. But their claims relating to PC take place in a liberal academic context and most of these authors use liberal normative concepts when justifying the censorship of particular speeches and actions (Delgado 1982; Kernohan 1998; Lawrence 1990; Levin 2010; Matsuda 1993).

4. Post-Modern Liberalism: Scientific Rationality as Political Incorrectness

While post-modern liberalism upholds the Enlightenment related values of individual liberty and equality, another Enlightenment value – that of autonomy reached by reason and pursuit of knowledge – fell by the wayside.

As Immanuel Kant admonished in his 1784 essay An Answer to the Question: What Is Enlightenment?: 'Sapere aude! Have the courage to use your own understanding! is thus the motto of enlightenment' (Schmidt 1996, p. 58). He called for the enlightened individual to 'dare to know', to use reason in order to disenthrall itself from immaturity. John Stuart Mill also asserted that the autonomous individual 'must use observation to see, reasoning and judgment to foresee, activity to gather materials for decision, discrimination to decide, and when he has decided, firmness and self-control to hold his deliberate decision' (Ten 2008, p. 47).

However, Western liberal societies that impose speech codes, prosecute microaggressions and ban speakers with controversial opinions from university campuses do not fit the picture of this described Enlightenment ideal of critical discourse. Hence, the question comes up, why did the Enlightenment values of reason and scientific rationality lose their importance in post-modern liberalism while other Enlightenment-related values, such as (individual) freedom and equality are still being held up? Joanna Williams offers a possible explanation when stating that, after the experience of the Holocaust during World War II, the Enlightenment promoted value of reason and its respective methods (rationality, the search for truth and empirical evidence) plunged into crisis: 'The Holocaust was considered by many to be a logical consequence of the endeavour to shape society through science and rationality' (Williams 2016, p. 63). Science as 'the emancipation of reason from emotions, of rationality from normative pressures, of effectiveness from ethics' (Bauman 1989, p. 108) came out of World War II as a failure and a succour of the Holocaust perpetrators. For those liberals disillusioned by scientific progress, post-modern liberalism became a viable option. Conversely, those others who saw war events as a product of irrationality can stand by enlightened liberalism.

A certain disappointment regarding the desirability of science had a particular consequence. Namely, truth-claims and the vision that a particular body of knowledge should assist us in moving closer to the truth became disreputable within parts of academia, especially in the radical humanities disciplines. As a result, some insights of critical post-modernism such as truth being relative and multiple replaced enlightened rationalism. Critical theory, developed by scholars from the Frankfurt School and later carried on by post-modernists like Michel Foucault, often questioned that to pursue knowledge and rationality would simply lead to truth-claims. Instead, they pointed to the seductive power of images and words, which these scholars perceive as having the potential to shape reality and to harm people (Williams 2016, p. 133). In this sense, truth-claims would rather implement and reinforce pre-existing power structures in society. According to Max Horkheimer and Theodor Adorno, 'technical rationality today is the rationality of domination. It is the compulsive character of a society alienated from itself (Horkheimer and Adorno 2002, p. 95). A critical and scientific discourse based on empirical evidence is then a tool of a political and economic power elite to strengthen its own position. In the words of Foucault:

> Truth is to be understood as a system of ordered procedures for the production, regulation, distribution, circulation and operation of statements. 'Truth' is linked in a circular relation with systems of power which produce and sustain it, and to effects of power which it induces and which extend it (Foucault 1980, p. 133).

As a result, this vision of truth and science 'undermines the ability to generate criteria for making ethical and political judgments, thereby threatening to plunge critical theory into relativism' (Bronner 2011, p. 33).

To be sure, not all critical theorists embrace post-modernism's incredulity towards universal scientific truth. For instance, Jürgen Habermas is a notorious critic of postmodern theory (Aylesworth 2015).

Yet, since its inception, critical theory emphasised how scientific and technological advancements are an instrument of domination in social relations (Horkheimer and Adorno 2002). Recent post-modern critical theory took one more step in this domination-oriented reasoning by casting out non-contingent scientific truth altogether. By doing so, critical post-modernism curtails the legitimacy of any potential governmental control undertaken in the name of objective truth. It is thus important to understand how bringing up epistemic relativism impacts the debate on PC.

First, there is not a single truth: Universities teach and uphold competing hypotheses. Still, there seem to be reservations towards making assertions that claim to be better and truer than other competing contentions. For example, while some feminist scholars (Grosz 1994; MacKinnon 1989; Prokhovnik 1999) claim physical differences between men and women (i.e. sex) are not responsible for behavioural differences (i.e. gender), there is a consensus among biologists, physicians and evolutionary psychologists that gender is (also) determined by biology (Baron-Cohen et al. 2005; Buss 1995; Hines 1982). However, these two competing assertions are both acknowledged within academia and are being taught on campus and published in leading international journals. Competing hypotheses within academia are the standard, but accepting no common standard of evaluation is likely to lead to parallel worlds of knowledge. Such worlds cannot assess one another without potential accusations of illegitimate authoritarianism.

Second, truth depends on perspective. The notion that knowledge is subjective leads way to contemporary identity politics. If truth is a personal construct, a heterosexual person, for instance, perceives the world in an entirely different way than a homosexual person. As a result, there cannot be a critical discourse about the accuracy of these two perspectives. None of them is truer than the other but they offer rather a distinctive point of view. According to some (Sue 2010; Waldron 2012), words have the potential to damage individuals at the psychological level; so to spread knowledge that historically disadvantaged groups and minorities may perceive as offensive is an act of aggression to avoid. Thus, it is not pc to claim certain knowledge is more valuable than another or to disconnect truth-claims from identity.

On the whole, contemporary post-modern liberalism has shifted away from the Enlightenment ideals of reason and scientific rationality. We may infer from the liberal egalitarian motivations behind this shift that liberalism neglected the value of striving for truth through knowledge and logic to protect itself from destruction via illiberal forces. As Michael Freeden notes, 'liberalism adapts through internal changes in the prioritization of its core concepts' (Freeden 2008, p. 15). And it seems adaptation was in order. What if because of a rational and scientific discourse someone established that individual freedom and universal equality are deficient ideas to construct the social order and that hierarchy and authority are systems which lead populations to greater success and satisfaction? By discrediting (objective) knowledge and critical reasoning, it is possible to diminish the potential danger of rational discourse for liberal tenets. In this sense, post-modern liberalism (Rorty 1992) seems to work as a purification of Enlightenmentliberal ideals, as an already tested and thus more robust version, which upholds certain liberal values, such as individual freedom and equality and therefore has to sacrifice idiosyncratic Enlightenment values, such as rationality and objective knowledge.

5. Liberal Science: A Veiled PC

As a reaction against post-modern PC advocates, Enlightenment liberals arose as the main opposing force to PC within socio-academic discourse. Although appearing to be fighting PC, this intellectual force ends up enforcing another version of the same phenomenon. Namely, they uphold science, reason and critical discourse but make sure potential illiberal findings or claims remain irrelevant. Enlightenment liberals defend liberal science[2] against PC because they believe identity-based thought control endangers liberalism. In particular, these writers claim PC is authoritarianism – especially speech restriction –, which endangers liberalism in its most dominant appearances: liberal democracy and liberal science (Chait 2015; Green 2006; Rauch 2013).

In his work *Kindly inquisitors – The new attacks on free thought*, Jonathan Rauch describes the liberal intellectual system (liberal science) as the only alternative to authoritarian orders (Rauch 2013, p. 28). Notably, Rauch shows two ways to rescue liberalism by reintroducing the Enlightenment ideal of reason and critical discourse.

First, Rauch asks for de-relativising knowledge. That is, to let liberal science decide about correct hypotheses (i.e. having knowledge) and incorrect claims (i.e. just having an opinion): 'Checking of each by each through public criticism is the only legitimate way to decide who is right' (Rauch 2013, p. 6). Hence, Rauch criticises the egalitarian attempt to relativise knowledge by respecting multiple truths and claims researchers should detect truth via critical discourse within liberal science.

Second, Rauch objurgates what he calls the humanitarian threat (Rauch 2013, p. 111) by asserting that the possibility of critical discourse is of higher importance to liberalism than the harm that offensive truth-claims can do to disadvantaged/minority groups. In order for liberal science to identify real knowledge, it cannot be 'nice (…). It does not give

[2] Liberal science is a term developed by Jonathan Rauch (2013) that represents an Enlightenment liberal intellectual system of knowledge production. It works with the following rules: no argument is really over; anyone can take part in scientific discussions. This system of knowledge production relies on the primacy of evidence and open discourse. As a term, liberal science remains in use, often by those opposing PC (Bailey 2005; J. Haidt and Lukianoff 2017).

a damn about your feelings and happily tramples them in the name of finding truth' (Rauch 2013, p. 19).

Hence, for Rauch, liberal science is the best mechanism to protect a liberal society from authoritarian measures. If everybody enjoys free speech and can put out truth-claims, the diverse scientific community sorts out the facts and disregards the errors. In this way, it is possible to avoid authoritarian decision makers who determine what is right and what is wrong: 'In an imperfect world, the best insurance we have against truth's being politicized is to put no one in particular in charge of it' (Rauch 2013, p. 110). On the one hand, according to Rauch, liberal science respects freedom of speech and belief; on the other hand, liberal science does not accept the right of beliefs to become knowledge straight away. Everybody can make claims all the time, but in order for claims to achieve the status of knowledge, they have to pass the process of the 'science game for checking' (Rauch 2013, p. 116). The idea here is to avoid empowering a political elite who then decides if something is knowledge or not. Instead, a competent but also diffuse scientific community (with no special interest in claiming power) controls the process of knowledge verification.

It is at least doubtful if liberals, such as Rauch, obey their strict rules of scientific discourse. Regarding potentially offensive truth-claims, Rauch (2013, p. 129) suggests ignoring offensive beliefs when they are uncontested or if liberal science already showed them to be wrong. In the same way as the post-modernists, liberal scientists may fear that through reason one may conclude that a liberal polity is undesirable. In fact, rational discussions within the scientific community often marginalise truth-claims whose implications question contemporary liberal morality. For instance, Duarte et al. show that liberals embed their values into investigation fields and methods. As a result, these liberals keep other researchers away from 'politically unpalatable research topics (…): areas such as race, gender, stereotyping, environmentalism, power, and inequality' (Duarte et al. 2015, pp. 1–2). So, the liberal scientific community is more likely to ignore or marginalise illiberal claims. Yet, to advocate free speech does not imply to refuse PC. Just because a scientist may enjoy free speech, it does not mean he can expect his controversial work to receive an objective and rational feedback within liberal science. Science is far from being self-correcting in matters of moral and political sensibility when there is an overarching moral consensus (Cofnas 2016), as it is the case with liberalism (Klein and Stern 2005). PC measures, such as pushing academics to liberal conformism, protect liberal hegemony.

First, Rauch illustrates the push for liberal conformism when stating that one should criticise or ignore hurtful opinions (Rauch 2013, p. 159). He is obviously supportive of neglecting controversial – assumable illiberal – truth-claims instead of engaging with difficult issues. This is a common position (Horgan 2013; Klein 2017; McWhorter 2017; Rose 2009). For instance, political theorist Steven Klein argues that we should allow individuals to present controversial (illiberal) truth-claims, but we should prevent them from entering the academic debate. As he puts it:

> Today, we've conflated a right to speak with a right to be taken seriously and debated. But while the former is a right, the latter is a privilege, and one that should be reserved for ideas that do not fundamentally threaten the foundations of our free and democratic society (Klein 2017).

Also, Steven Pinker, who notably criticises the damaging effects of PC on social and scientific discourse, opens specific exceptions for the 'benign taboos on racism, sexism and homophobia' (Pinker 2018, p. 219). He clarifies that we should be 'mindful of excessive taboos' because they can diminish the credibility of journalists and academics (Pinker 2018), yet he is not claiming we should be mindful of – liberal – taboos per se. This overall ethical approach can be partly responsible for young scholars avoiding controversial areas of research as it contributes to a climate of liberal conformity among academics.

Second, Rauch's claim that we should not try 'to silence or punish' people who hold discriminatory opinions but instead try 'to correct them' (Rauch 2013, p. 181) implies that truth-claims with discriminatory content are (morally) wrong and therefore we must amend them. Likewise, Pinker argues that academic free speech is necessary because freedom of expression allows us to use rationality to put controversial facts in a liberal context, which helps to avoid illiberal dangerous conclusions (Pinker 2017). Apparently, thinkers like Pinker and Rauch are self-assured that reason will never give support to non-liberal forms of political organisation. By this means, they show that they do not understand science as a process with an open outcome but as a process whose duty is to protect liberalism. There is also another stated reason for why potentially offensive speech should be permissible: 'And what about the day when right-wingers get the upper hand? Will they be fair?' (Rauch 2013, p. 143). It exists a latent fear that the 'inquisition' (Rauch 2013, p. 27) put in place by egalitarians and humanitarians to defend their vision of freedom and equality leads to authoritarian structures which an up-coming inegalitarian regime may use. A central aim of liberal science is to prevent illiberal political power from arising.

Enlightenment liberals advocate free speech and support the de-relativisation of knowledge. However, if their critical discourse only engages with claims and theories that remain within the liberal framework, if they ignore or marginalise claims outside this framework, they endorse a different kind of PC. Specifically, a PC that does not act authoritarian by forbidding offensive expressions and filing anti-discrimination laws but a PC that rather pushes people to perform self-censored conformist behaviour in order not to get marginalised. Liberal science worries that an authoritarian and identitydriven PC, as carried out by egalitarians and humanitarians, harbours the danger of triggering an illiberal identitarian counter-movement. As social psychologist Jonathan Haidt explains it:

> If you keep treating white men as an identity group, you keep saying that 'they are terrible; they are evil' – eventually they become just like another identity group and they vote[d] their racial interests, in a sense you might say. So identity politics on the Left eventually triggers identity politics on the Right (Jonathan Haidt 2016).

Likewise, in defence of liberal science and moral individualism, the prominent anti-PC activist and psychologist Jordan Peterson clarifies that both identity politics – from the Left and from the Right – are 'equally dangerous' (Luscombe 2018). Thus, the tactic of ignoring, marginalising and not offering critical engagement with system-challenging opinions relies on the central goal of preventing the rise of identitarian illiberalism (Pinker 2018, p. 143; Rauch 2013). This goal and result oriented science promoted by Enlightenment liberals does not seem to have much in common with Kant's 'dare to know' attitude towards science. Instead, it bears similarities to Karl Popper's (1945) advocacy of intolerance towards illiberal discourses as the best way to protect the open society.

All in all, it is possible to understand that both the post-modern advocates of PC and their Enlightenment liberal opponents make up two sides of the same coin. On the side of post-modern PC, traditional Enlightenment values of reason and rationality got partly ejected from contemporary liberalism, being replaced by relativism and perspectivism. On the side of the Enlightenment liberals, there seems to exist liberal truth-claims that they do not debate and take for granted; so they marginalise or ignore claims challenging these pre-assumed positions. In this context, Williams asserts that Enlightenment liberals often assume that:

> the truth of a particular issue is settled beyond question. The tendency to label critics, or skeptics, on issues as wide ranging as the Holocaust, climate change, patriarchy and rape culture, as 'deniers' suggests not a clash of opposing understandings but that the truth has already been determined and people who do not accept it are deluded. It suggests that any further discussion is not only futile but problematic as it detracts from dealing practically with the issues concerned (Williams 2016, p. 67).

Williams understands that 'both the rejection of truth and the notion that the truth is settled curtail academic debate by undermining the assumption that knowledge progresses through competing truth claims' (Williams 2016, p. 67).

6. PC and the Disconnection between Liberalism and the Enlightenment

The moral positions that sprung from Enlightenment thought are not uniform. Particularly at the moral or ideological level, we can speak of Enlightenments, plural. Yet, Enlightenment liberals conflate liberalism and the Enlightenment as if these two concepts were interchangeable. The two concepts represent in fact two different traditions. As the likes of Nietzsche (2009) and Tocqueville (1959) realise, liberalism's defence of liberty and equality in universalistic and individualistic terms derives from Christian monotheism. In contrast, the Enlightenment defence of reason and scientific rationality evolved from ancient Greek thought, which often operated in a (pagan) nonliberal moral framework. Both Aristotelian and Platonic streams of thought were deeply biopolitical and strongly concerned with controlling the quality of population, therefore deriving moral worth from a hierarchical biological status (Ojakangas 2016). In this sense, to uphold scientific rationality does not require liberalism.

Without doubt, Enlightenment thinkers were not all liberal. Most notably, Auguste Comte's rejection of Christian-liberal metaphysics (e.g. human rights) led him to advocate a new 'religion of humanity', where scientific experts of the industry would discover the most appropriate moral framework for society (Comte 1927). As John Gray points out, 'the link between the Enlightenment and liberal values (…) is actually rather tenuous. It is strongest in Enlightenment thinkers who were wedded to monotheism, such as Locke and indeed Kant' (Gray 2018). Those unwedded to monotheism oftentimes espouse non-liberal values informed by science (Ojakangas 2016).

The close relationship between the Enlightenment advocacy of science and illiberalism is now an influential idea in scholarly terms, especially among critical perspectives (Geuss 1998). In particular, Adorno and Horkheimer's The Dialectic of Enlightenment (2002) disseminated the tight link between science and illiberalism. In this book, the two authors focus on the social consequences of instrumental reason, which is the capacity to discover effective means to satisfy whatever ends an agent may have. In its most sophisticated form, instrumental reason aims at finding scientific truth while remaining morally agnostic. They think the findings of empirical science alone cannot validate Enlightenment liberal ideals. For them, if facts are the single source of knowledge, 'in the end the (liberal) ideals themselves come to look like myths or prejudices which ought to be discarded' (Geuss 1998), thus opening the way to an explicit dominance hierarchy. However, the scientific knowledge of the natural world may be capable of identifying objective values — a standard philosophical position within natural moral realism (Richards 2017). But whether or not science can identify true moral values, Adorno and Horkheimer understand that Enlightenment liberal values are not free from naturalist scrutiny.

To embrace scientific rationality altogether – by removing it from unnaturalistic metaphysics – should mean that one is open to revising moral values according to the progress of knowledge. By making a case against PC and in favour of critical discourse, Enlightenment liberals should be open to moral revision. After all, morality is a social phenomenon thoroughly studied by science (Ruse and Richards 2017). Still, they do not show the willingness to revise their values according to science and continue to understand liberalism as having priority over scientific reason. For instance, Pinker claims scientific reason justifies liberal cosmopolitanism and disproves the value of ingroup favouritism. He asserts that

> reason goads us into realizing that there can be nothing uniquely deserving about ourselves or any of the groups to which we belong. We are forced into cosmopolitanism: accepting our citizenship in the world (Pinker 2018, p. 11).

Yet, numerous scientific theorists demonstrate the importance of in-group favouritism in the evolutionary system (Axelrod and Hammond 2006; Faria 2017; Hartshorn et al. 2013), making his normative claim far from scientifically informed. The assertion that science only validates liberal values is another form of PC, which delegitimises illiberal scientific claims within the academic sphere.

Perhaps all ideological positions defend a set of values that demarcate no-go areas of belief, and liberalism is no exception. But one should not confuse the defence of values with PC. If so, all defences of certain value preferences over others would make up PC. Instead, the pc phenomenon is about silencing modes of expression antagonistic to one's value systems.[3] What Enlightenment liberals do is not only to defend liberal values through reasoned rhetoric. Contrary to their open discourse narrative, they have an active role in silencing dissident voices by dismissing or overlooking illiberal truth-claims. Although outright censorship is not part of liberal science, this worldview engages in subtle mechanisms that restrict undesired positions.

For instance, liberal science conflates (liberal) value and (scientific) fact. By theoretically promoting critical discourse and the quest for truth, Enlightenment liberals seem to believe that no truth can ever justify illiberal claims. They seem to think truth has already validated liberalism, therefore making it safe to ignore illiberal truth-claims, portraying these claims as the product of bad science (Barber 2013; Newby and Newby 1995). Pinker (2017) goes so far as saying that any empirical truth should appear in critical academic forums for academics to diffuse potential illiberal truth-claims. Likewise, self-described conservative author Andrew Sullivan claims dangerous findings on genetic group differences obliges us to

> establish a liberalism that is immune to such genetic revelations, that can strive for equality of opportunity, and can affirm the moral and civic equality of every human being on the planet (Sullivan 2018).

Enlightenment liberals set two levels of scientific truth-seeking, one open to most facts and another one closed to non-liberal moral claims deriving from threatening facts.

The scientific behavioural fields that study average genetic differences between social groups regarding race and sex provide several examples of how PC became a method to protect liberal values. For instance, Noam Chomsky notes that these studies are of 'no scientific interest and of no social significance, except to racists, sexists, and the like'

[3] Unlike other ideologies that do not sacralise free expression, liberalism has a special inner tension because freedom of expression is an important part of the traditional liberal ethos.

(Chomsky 1988, p. 164). Distinguished psychologist Howard Gardner calls the researchers that find natural inequalities between groups 'bad guys' (Gardner 2009) and 'pseudo-scientists' (Gardner 2001, pp. 6–7). In turn, evolutionary philosopher Daniel Dennett reveals the classic strategy to protect liberal morality from dangerous findings that reveal natural inequalities between groups (e.g. IQ, personality traits):

> if I encountered people conveying a message I thought was so dangerous that I could not risk giving it a fair hearing, I would be at least strongly tempted to misrepresent it, to caricature it for the public good. I'd want to make up some good epithets, such as genetic determinist or reductionist or Darwinian Fundamentalist, and then flail those straw men as hard as I could. As the saying goes, it's a dirty job, but somebody's got to do it (Dennett 2003, pp. 19-20).

Dennett seems to claim that these dangerous findings should remain outside of scientific discourse regardless of how good the evidence may be. Yet, he made no claims of legally forbidding research, which reminds us that PC, in the form of softcensorship, often operates without outright prohibition and instead manifests itself through social pressure towards conformity (Loury 1994, p. 430). On another occasion, Dennett (2006, p. 337) condemns lying about scientific facts when other political forces like Marxism do it, showing that the defence of liberalism justifies the means.

After surveying the scientific community's attitudes towards these 'dangerous' topics since the 1970s, Nathan Cofnas concludes that within the community there is a widespread acceptance of two central ideas:

(1) the prevailing morality requires that 'scientists should not conduct research that threatens to uncover facts that contradict these morally required beliefs' and
(2) the same morality 'requires people to hold certain beliefs regardless of the evidence' (Cofnas 2016, p. 479).

Given these widespread beliefs, it is not surprising that Enlightenment liberals are merely engaged in another form of PC, which, although rejecting the post-modern disregard for the truth, equally disregards the importance of knowledge when it conflicts with liberal values.

Ultimately, both post-modern advocates of PC and Enlightenment liberals deliberately conflate fact and value. The former understand scientific knowledge not as truth but as a narrative of power, while the latter are interested in truth as long as it validates liberalism as objectively good. Remarkably, no side seems to believe in the strict separation of is from ought, which is clear in the shared fear that the discovery of empirical facts can lead to illiberal normative claims. As philosopher Robert J. Richards showed in his defence of evolutionary ethics, the reason it is so complicated to separate is from ought is that moral justification

> must ultimately lead to an appeal to the beliefs and practices of men, which of course is an empirical appeal. So moral principles ultimately can be justified only by facts (Richards 1986, p. 286).

Because of this prevalent conflation of fact and value, the socio-academic debate about PC is actually a debate about how to better protect liberalism. Strikingly, there is a general absence of critiques of PC that are truly open-ended regarding (scientific) truth and its moral consequences, including those of a potentially illiberal nature. A possible explanation for this absence has two dimensions:

(1) the academic community is overwhelmingly liberal, leading to a general lack of moral diversity and to a weak pluralism (D. B. Klein and Stern 2005).
(2) PC itself, with its soft penalties at the social and professional levels, makes up a barrier against the existence of open-ended critiques of PC.

This leads us to a full circle where the debates about PC are themselves pc.

7. Conclusion

The socio-academic debate about PC presents a dichotomy between critical postmodern advocates of PC and those Enlightenment liberals who oppose PC. Yet, we showed that this dichotomy does not hold under scrutiny and that both sides are ultimately defenders of PC who merely use different pc strategies. In particular, Enlightenment liberals represent a concealed form of PC. Both sides are more interested in defending liberal values than in unfettered critical discourse. While postmodern advocates of PC straightforwardly dismiss objective truth, Enlightenment liberals uphold the existence and the desirability of truth. Yet, these science-based liberals are in fact protecting liberalism from an uncompromising open-ended quest for scientific truth.

As demonstrated, both sides use PC because PC works as a mechanism to protect and further liberal values. Its central aim is to prevent the rise of illiberal truth-claims. Hence, the socio-academic debate about PC is not a debate between two factions in favour and against PC, but a debate about how to better protect and further liberalism. It is a debate about the kind and degree of PC restrictions that can best defend liberalism from illiberal truth-claims and political stances. On one side, post-modern advocates wish to censor political incorrectness due to their understanding of some truth-claims as narratives of oppression. These advocates aim to suppress such narratives in the name of liberating tolerance. On the other side, Enlightenment liberals are more inclined to marginalise dangerous scientific research. Although falling short from banning dangerous speech, they reject the moral and scientific legitimacy of truth-claims that fall outside of the liberal paradigm. Enlightenment liberals often assert that certain (liberal) truth-claims are scientifically sound and beyond sensible debate.

Last, by noting that the Enlightenment differs from liberalism, we argued that Enlightenment principles of truth-seeking and critical discourse may also operate in non-liberal moral spheres or lead to them. Hence, to conflate liberalism and Enlightenment reveals not a commitment to open-ended scientific rationality but, above all, a commitment to liberalism and its (PC) safeguard. The debate on PC lacks a prominent anti-PC side arguing for an open-ended critical discourse at the scientific and moral levels, an absence likely caused by liberal hegemonic thought in academia and by PC itself. As a result, the PC debate represents a circular and closed dispute about how to uphold liberal values.

References

Axelrod, R., & Hammond, R. A. (2006). The evolution of ethnocentrism. Journal of Conflict Resolution, 50(6), 926–936.

Aylesworth, G. (2015). Postmodernism. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy.

Bailey, R. (2005). The triumph of liberal science. Reason. Retrieved from http://reason.com/archives/2005/12 /30/the-triumph-of-liberal-science

Barber, A. (2013). Science's immunity to moral refutation. Australasian Journal of Philosophy, 91(4), 633–653.

Baron-Cohen, S., Knickmeyer, R. C., & Belmonte, M. K. (2005). Sex differences in the brain: Implications for Explaning autism. Science, 310(5749), 819–823.

Bauman, Z. (1989). Modernity and the holocaust. Cambridge: Polity Press.

Bernstein, R. (1990, 28 October). The rising hegemony of the politically correct. The New York Times.

Bernstein, D. E. (2003). You can't say that!: The growing threat to civil liberties from antidiscrimination laws. Washington, D.C.: Cato Institute.

Bloom, A. D. (1987). The closing of the American mind: How higher education has failed democracy and impoverished the souls of today's students. New York: Simon & Schuster.

Brink, B. v. d. (2000). The tragedy of liberalism: An alternative defense of a political tradition. Albany: State University of New York Press.

Bronner, S. E. (2011). Critical theory: A very short introduction Retrieved from http://www. veryshortintroductions.com/view/10.1093/actrade/9780199730070.001.0001/actrade-9780199730070

Buss, D. M. (1995). Psychological sex differences: Origins through sexual selection. American Psychologist, 50(3), 164–168.

Byrne, J. M. (1997). Religion and the Enlightenment: From Descartes to Kant. Louisville: Westminster John Knox Press.

Chait, J. (2015, 26 January). Not a Very P.C. Thing to Say: How the language police are perverting liberalism. New York Magazine.

Chomsky, N. (1988). Language and problems of knowledge. The Managua lectures. Cambridge: MIT Press.

Cofnas, N. (2016). Science is not always self-correcting. Foundations of Science, 21(3), 477–492.

Cole, J. R. (2006). Academic freedom under fire. Daedalus, 134(2), 5–17.

Comte, A. (1927). Sociology and the new politics. American Journal of Sociology, 33(3), 371–381.

Cook, P., & Heilmann, C. (2012). Two types of self-censorship: Public and private. Political Studies, 61(1), 178–196.

Delgado, R. (1982). Words that wound: A tort action for racial insults, epithets, and name calling. Harvard Civil Rights-Civil Liberties Law Review, 17, 133.

Dennett, D. (2003). Freedom evolves. New York: Viking.

Dennett, D. (2006). Breaking the spell: Religion as a natural phenomenon. New York: Viking.

Dryzek, J. S. (2000). Deliberative democracy and beyond. Liberals, critics, contestations. Oxford: Oxford University Press.

D'Souza, D. (1991). Illiberal education: The politics of race and sex on campus. New York: Maxwell Macmillan International.

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. E. (2015). Political diversity will improve social psychological science. Behavioral and Brain Sciences, 38(130), 1–58.

Eddy, M. D. (2017). The politics of cognition: Liberalism and the evolutionary origins of victorian education. British Journal for the History of Science, 50, 677–699.

Faria, F. N. (2017). Is market liberalism adaptive? Rethinking F.a. Hayek on moral evolution. Journal of Bioeconomics, 19(3), 307–326.

Feldstein, R. (1997). Political correctness: A response from the cultural left. Minneapolis: University of Minnesota.

Fish, S. E. (1994). There's no such thing as free speech, and it's a good thing too. New York: Oxford University Press.

Foucault, M. (1980). Power/knowledge: Selected interviews and other writings, 1972–1977. New York: Vintage.

Freeden, M. (2008). European liberalism. An essay in comparative political thought. European Journal of Political Theory, 7(1), 9–30.

Furedi, F. (2016). What's happened to the University?: A sociological exploration of its infantilisation.

Gardner, H. (2001). The ethical responsibilities of professionals. In J. Solomon (Ed.), The good work project. Retrieved from http://thegoodproject.org/pdf/2-Ethical-Resp-of-Prof-7_98.pdf.

Gardner, H. (2009). Intelligence: It's not just IQ. The Rockefeller University: Parents & Science. Retrieved from https://www.youtube.com/watch?v=ESGLRnitp4k

Geuss, R. (1998). Critical theory. Routledge Encyclopedia of Philosophy. Retrieved from https://www.rep. routledge.com/articles/thematic/critical-theory/v-1

Gitlin, T. (1995). The twilight of common dreams: Why America is wracked by culture wars (1st ed.). New York: Metropolitan Books.

Gray, J. (2018). Unenlightened thinking: Steven Pinker's embarrassing new book is a feeble sermon for rattled liberals. New Statesman. Retrieved from https://www.newstatesman.com/culture/books/2018/02 / unenlightened-thinking-steven-pinker-s-embarrassing-new-book-feeble-sermon

Green, D. G. (2006). We're (nearly) all victims now! How political correctness is undermining our liberal culture. London: Civitas.

Grosz, E. A. (1994). Volatile bodies: Toward a corporeal feminism. Bloomington: Indiana University Press.

Haidt, J. (2016) The psychology behind rising nationalism. BBC Newshour.

Haidt, J., & Lukianoff, G. (2017). Why it's a bad idea to tell students words are violence. The Atlantic. Retrieved from https://www.theatlantic.com/education/archive/2017/07/why-its-a-bad-idea-to-tellstudents-words-are-violence/533970/

Hartshorn, M., Kaznatcheev, A., & Shultz, T. (2013). The evolutionary dominance of ethnocentric cooperation. Journal of Artificial Societies and Social Simulation, 16(3), 7.

Hildebrandt, M. (2005). Multikulturalismus und Political Correctness in den USA (1. Aufl. ed.). Wiesbaden: VS Verl. für Sozialwiss.

Hines, M. (1982). Prenatal gonadal hormones and sex differences in human behavior. Psychological Bulletin, 92(1), 56–80.

Hollander, P. (1994). "Imagined tyranny"? Political correctness reconsidered. Academic Questions, 7(4), 51–73.

Horgan, J. (2013, 16 May). Should research on race and IQ be banned? Scientific American. Retrieved from https://blogs.scientificamerican.com/cross-check/should-research-on-race-and-iq-be-banned/

Horkheimer, M., & Adorno, T. W. (2002). Dialectic of enlightenment: Philosophical fragments. Stanford: Stanford University Press.

Kernohan, A. W. (1998). Liberalism, equality, and cultural oppression. New York: Cambridge University Press. Kimball, R. (1990). Tenured radicals: How politics has corrupted our higher education. London: Harper & Row.

Klein, S. (2017, 6 November). Right-wing extremists have a right to speak, not a right to be listened to. The Washington Post. Retrieved from https://www.washingtonpost.com/news/posteverything/wp/2017/11/06
/right-wing-extremists-have-a-right-to-speak-not-a-right-to-be-listened-to/?utm_term=.792e263d2ade

Klein, D. B., & Stern, C. (2005). Professors and their politics: The policy views of social scientists. Critical Review, 17(3–4), 257–303.

Knowles, E., & Elliott, J. (1997). The Oxford dictionary of new words (new ed.). Oxford: Oxford University Press.

Kors, A. C., & Silverglate, H. A. (1998). The shadow university: The betrayal of liberty on America's campuses. New York: Free Press.

Lawrence, C. R. (1990). If he hollers let him go: Regulation racist speech on campus. Duke Law Journal, 1990, 431–437.

Lea, J. (2009). Political correctness and highereducation: British and American perspectives. London: Routledge. Levin, A. (2010). The cost of free speech: Pornography, hate speech and their challenge to liberalism. Basingstoke: Palgrave Macmillan.

Locke, J. (1988). Two treatises of government (student ed.). Cambridge Cambridgeshire. New York: Cambridge University Press.

Loury, G. C. (1994). Self-censorship in public discourse. A theory of 'political correctness' and related phenomena. Rationality and Society, 6(4), 428–461.

Lukianoff, G., & Haidt, J. (2015, September). The coddling of the American mind. The Atlantic. Retrieved from https://www.theatlantic.com/magazine/archive/2015/09/the-coddling-of-the-american-mind/399356/ Luscombe, B. (2018). Jordan Peterson talks gun control, angry men and why so few women lead companies.

Time. Retrieved from http://time.com/5175974/jordan-peterson-12-rules-book-interview/

MacKinnon, C. A. (1989). Toward a feminist theory of the state. London: Harvard University Press.

Marcuse, H. (1965). Repressive tolerance. In R. P. Wolff & B. Moore (Eds.), A critique of pure tolerance (pp. 81–117). Boston: Beacon Press.

Matsuda, M. (1989). Public response to racist speech: Considering the victim's story. Michigan Law Review, 87(8), 2320–2381.

Matsuda, M. (1993). Words that wound: Critical race theory, assaultive speech, and the first amendment. Boulder: Westview Press.

McWhorter, J. (2017). Stop obsessing over race and IQ. National Review. Retrieved from https://www. nationalreview.com/2017/07/race-iq-debate-serves-no-purpose/

Moller, D. (2016). Dilemmas of political correctness. Journal of Practical Ethics, 4(1), 1–22.

Newby, R. G., & Newby, D. E. (1995). The bell curve: Another chapter in the continuing political economy of racism. American Behavioral Scientist, 39(1), 12–24.

Nietzsche, F. (2009). On the genealogy of morals. A polemical tract (I. Johnston, Trans.). Arlington, Virginia: Richer Resources Publications.

Ojakangas, M. (2016). On the greek origins of biopolitics: A reinterpretation of the history of biopower. Oxon: Routledge.

Parekh, B. (2017). Limits of free speech. Philosophia, 45(3), 931–935.

Pinker, S. (2018). Enlightenment now: A manifesto for science, reason, humanism, and progress. New York: Penguin.

Pinker, S. (2018). Going rogue: Political correctness. World Economic Forum. Retrieved from https://www.youtube.com/watch?v=fFohRupaXzc

Pinker, S. (2017). Is political correctness why trump won?, Spiked Magazine Panel.

Popper, K. R. (1945). The open society and its enemies. London: Routledge & Kegan Paul.

Prokhovnik, R. (1999). Rational woman: A feminist critique of dichotomy. London: Routledge.

Rauch, J. (2013). Kindly inquisitors: The new attacks on free thought (expanded edition. ed.). Chicago: The University of Chicago Press.

Rawls, J. (1993). Political liberalism. New York: Columbia University Press.

Richards, R. J. (1986). A defense of evolutionary ethics. Biology and Philosophy, 1, 265–293.

Richards, R. J. (2017). Evolutionary ethics: A theory of moral realism. In R. J. Richards & M. Ruse (Eds.), The Cambridge handbook of evolutionary ethics (pp. 143–157). Cambridge: Cambridge University Press.

Rorty, R. (1992). The priority of democracy to philosophy. In J. P. Reeder & G. Outka (Eds.), The priority of democracy to philosophy (pp. 254–278). Princeton: Princeton University Press.

Rorty, R. (1998). Achieving our country: Leftist thought in twentieth-century America. Cambridge: Harvard University Press.

Rose, S. (2009). Should scientists study race and IQ? NO: Science and society do not benefit. Nature, 457, 786–788.

Roxburgh, A. (2002). Preachers of hate: The rise of the far far right. London: Gibson Square Books.

Ruse, M., & Richards, R. J. (2017). The Cambridge handbook of evolutionary ethics. Cambridge: Cambridge University Press.

Schmidt, J. (1996). What is enlightenment?: Eighteenth-century answers and twentieth-century questions. Berkeley: University of California Press.

Sparrow, R. (2002). Talking sense about political correctness. Journal of Australian Studies, 26(73), 119–131. Sue, D. W. (2010). Microaggressions in everyday life: Race, gender, and sexual orientation. Hoboken: Wiley.

Sullivan, A. (2018). Denying genetics isn't shutting down racism, it's fueling it. New York magazine. Retrieved from http://nymag.com/intelligencer/2018/03/denying-genetics-isnt-shutting-down-racism-its-fueling-it.html Ten, C. L. (2008). Mill's on liberty: A critical guide. Cambridge: Cambridge University Press.

Tocqueville, A. d. (1959). The european revolution and correspondence with Gobineau. New York: Doubleday.

UDHR. (2010). The universal declaration of human rights, claiming human rights.

Waldron, J. (1993). Liberal rights: Collected papers, 1981–1991. Cambridge: Cambridge University Press.

Waldron, J. (2012, 20 March). The harm of hate speech. Free Speech Debate. Retrieved from http://freespeechdebate.com/discuss/the-harm-of-hate-speech/

Williams, J. (2016). Academic freedom in an age of conformity: Confronting the fear of knowledge. New York. Wilson, J. K. (1995). The myth of political correctness: The conservative attack on higher education. Durham: Duke University Press.

Zafirovski, M. (2011). The enlightenment and its effects on modern society. New York: Springer.

Individual Liberty and the Importance of the Concept of the People

Regina Queiroz

Queiroz, R. (2018). Individual liberty and the importance
of the concept of the people.
***Palgrave Communications, 4*, 1–42.**
https://doi.org/10.1057/s41599-018-0151-3

Introduction

Through publically agreed laws that correspond to a common set of public restrictions, the 'people as a sovereign body' serves to protect against violations of individual liberty and despotic power (Locke, 1679 (1960); Kant, 1793 (1977)). Where no such common body exists, individuals are deprived of this protection. In such cases, individuals must obey without liberty, while those in power command under a state of license, i.e., a state of unrestricted liberty. Neoliberal theorists maintain that any common personality, with its corresponding set of public restrictions on liberty, undermines individual liberty (Hayek, 1976; Nozick, 1974). Therefore, in addition to promoting the idea of private, atomized individuals and denying the existence of "the people" (Hayek, 1976; Nozick, 1974), neoliberal theory permits only private restrictions (positive and negative) on liberty (Hayek, 1976; Nozick, 1974).

Against this neoliberal assumption (Hayek, 1976; Nozick, 1974), we shall argue that rejecting the concept of the people and public restrictions on liberty while preserving the general law, its protective function, and coercive institutions and instruments for enforcing neoliberal law poses a serious threat to individual liberty and ultimately risks reducing the majority of free individuals to servile – and in some cases lawless – persons.

The literature has already demonstrated the incompatibility between neoliberalism and the notion of the people as a political category and reality (Brown, 2015; Dean, 2008). The impact of neoliberalism's exclusion of the people and its reliance on the concept of publicity without a public has also been demonstrated (Queiroz, 2017). Related to this, the literature has addressed how neoliberalism fosters the development of a docile and disciplined citizenry (Foucault, 2008). Nonetheless, the political consequences of the exclusion of the people and the protective role it plays in the preservation of the political state – namely the transformation of free individuals into servile, and ultimately lawless, persons – has yet to be addressed, in particular from a political-philosophical point of view.

The importance of this issue is clear. There has been much emphasis on the economic nature of neoliberalism, which has obscured the fact that, more than an economic position, neoliberalism is a political outlook and reality (Bruff, 2014). Although neoliberalism has become deeply tied to economics (Hall, 2011; Read, 2009), this is mainly due to the fact that its theoretical understanding of the state as a political institution is made in analogy with the economic market and the subsequent political redefinition of the latter's aims and scope (Foucault, 2008). Thus, without neglecting the significance of neoliberal economic analysis, in shifting the focus to neoliberalism's political character we aim to disclose its political-philosophical foundations and to translate its allegedly purely economic aspects to the political sphere. As we will see, the imposition of fiscal equilibrium, fiscal consolidation, cuts to social security, the privatization of public property, the liberalization of collective bargaining, and the shrinking of pensions (Barro, 2009) are connected not only to the rise of poverty and inequality but also to the transformation of free citizens into dependent and servile persons.

The underlying philosophical principles formulated in Hayek's political economy, political philosophy and legal theory, as well as in Nozick's libertarianism, have spilled over into politics. Although, as empirical studies frequently show, there is always a gap between theoretical statements and practical reality, these principles now provide, at a national and international level, the law's substantive content (Brown, 2015; Gill, 1998; Hall, 2011; Klein, 2007; Overbeek, 1993).

For these reasons, we do not intend to evaluate the "exegetical" value of Hayek's and Nozick's philosophical views (for example Hayek's mistaken reading of Kant's ethical and political philosophy; Gray, 1989). At the same time, we cannot here explore the important material basis of neoliberal ideology, namely concrete neoliberal activities, processes and powerful neoliberal social and political forces, such as multinational corporations (Brown, 2015; Gill, 1998; Hall, 2011; Harvey, 2005; Klein, 2007; Overbeek, 1993). Instead, we aim to show that the philosophical assumptions underlying Hayek's political economy and Nozick's libertarianism allow us to clarify the connection between the exclusion of the people as a political category and neoliberalism's promotion of a servile citizenry.

To better understand this connection, this paper will consider the Lockean and Kantian concepts of the people. Despite the differences between Locke's and Kant's political philosophies (Gray, 1989; Williams, 1994), for both thinkers the people serves the function of protecting individual liberty against despotic power, a condition which is commonly referred to as political obligation under liberty. Hayek and Nozick explicitly refer to the Lockean and Kantian foundations of their views, for example the Kantian universalization test for establishing the validity of the abstract rules of the market state (Hayek, 1976). Nozick's use of the Kantian understanding of the person as an end in itself to justify the rejection of substantive principles of justice (Nozick, 1974) provides an additional reason to consider Locke's and Kant's conceptions of the people in detail.

There are of course important differences between our current social, political and technological context, which is characterized by globalization, and Locke and Kant's modern nation states. We ought also to consider the differences between how we conceive of the people, e.g., whether we define peoples in terms of national commonality (Miller, 2000) or whether we ought to stress the role of democratic politics in creating this sense of political belonging (Habermas, 2008). Equally significant is the fact that, contrary to neoliberalism, Locke's liberalism depends on *homo politicus* and *juridicus* rather than *homo economicus*, which generates significant tensions between his rights-based view and modern views based on interests (Foucault, 2008). Equally, we wish to overlook neither Locke's and Kant's controversial statements and practices, for example Kant's exclusion of non-property-owners from the social contract (Kersting, 1992), nor the limits of Locke's and Kant's theoretical constructions of political personality (Badiou, 2016). The weaknesses of past democracies, expressed in the exclusion of woman from equal citizenship, the existence of slavery, and contemporary populist perversions of democracy, do not entail that we must abandon the ideal of democratic political power, however. The negative aspects of Locke's and Kant's political philosophies should not erase their strong commitment, from a liberal perspective, to the importance of the concept of the people when it comes to protecting individual liberty.

Finally, we do not wish to ignore past conceptions of the people, such as Greco-Roman conceptions, republican conceptions (Cicero, 1999; Habermas, 2000; Rousseau, 1762 (1964)), Marxist conceptions (Badiou, 2016), and other current alternatives. Despite their differences, they share certain features with the liberal approach, such as assigning a protective role to the people. In the face of the political consequences of neoliberalism's exclusion of the people, we should appeal to what Rawls (1993) calls overlapping consensus, i.e., agreement on the people as a political category on different grounds.

The paper is organized as follows. Section 1 provides a brief presentation of the main concepts and neoliberalism's rejection of public restrictions on liberty and the right to equal and reciprocal coercion. In the second section, we show that, contrary to neoliberal assumptions, far from fostering individual liberty, the exclusively private restriction of liberty implies a political distinction between those who obey and those who rule. It also entails the division of citizens into those who obey and those who command, where the latter are given unequal protection by the government and thus an unequal share in the public coercive power. Similarly, it involves the introduction of two familiar political categories, originally deployed in neoliberal political society: self-serfdom on the one hand and invisible, voiceless citizenship on the other. At the end of the paper, we provide a brief account of the protective role of the people as a political body when it comes to individual liberty. We show that by ensuring the equal and reciprocal right of coercion, the people as a body protects individual liberty.

The people vs. the private coercion of liberty under neoliberalism

As an imprecise and nebulous concept, there is no single "pure" form of neoliberalism. Instead, there are varied articulations that make up an extraordinarily messy amalgam of neoliberal ideas and policies at multiple sites (Latin America, Europe, China; Harvey, 2005), on multiple scales (national, international, transnational, global; Brown, 2015; Hall, 2011; Klein, 2007; Overbeek, 1993), and within the many versions of the welfare state (Kus, 2006). Additionally, according to England and Ward's (2016) taxonomy, neoliberalism can be thought of as a form of statecraft that promotes the reduction of government spending while increasing economic completion (Mudge, 2008), or as a form of governmentality that comprises social, cultural and economic practices that constitute new spaces and subjects (Foucault, 2008). In addition, neoliberalism can be seen as a reaction to the disenchantment identified by Weber, (1978) following the rise of bureaucracy. Neoliberalism expresses a kind of re-enchantment with the exclusively individual rational actor, who claims a nonalienable space of liberty against a bureaucratic "iron cage". Although some see neoliberalism as a privatized version of economic and bureaucratic despotism (Lorenz, 2012) or as a totalizing global bureaucracy (Hickel, 2016), this re-enchantment can explain the enthusiastic endorsement of neoliberal principles by a wide spectrum of political and ideological forces, for example by the Labour party under Blair in Great Britain, the SPD under Schröder in Germany, and followers of Pinochet in Chile.

Finally, neoliberalism has been viewed as a conception of the world, or a "total view of reality" (Ramey, 2015, p. 3), which is meant to be applied to the political realm and the entirety of human existence. Integrated into common sense, its main ideas stem from the everyday experience of buying and selling commodities on the market, a model that is then transferred to society. As a total view of reality, neoliberalism entails "a new understanding of human nature and social existence [and] the way in which human beings make themselves and are made subjects" (Read, 2009, p. 28; see also Foucault, 2008).

While acknowledging the disparate criteria for defining and assessing neoliberal theory and practice, we maintain that neoliberalism is a political outlook and reality (Bruff, 2014) which has evolved in part in accordance with the framework of the theoretical premises of Hayek's, (1976) political economy and Nozick's, (1974) philosophical libertarianism. For instance, neoliberal theoretical principles now provide, at a national and international level, substantive content to political constitutions (McCluskey, 2003), the establishment of laws governing the executive (Foucault, 2008; Read, 2009), and the reformulation of laws governing citizens (LeBaron, 2008; McCluskey, 2003; Supiot, 2013, p.

141; Wacquant, 1999). They also shape our comprehension of the world and ourselves (for example the reduction of the citizen to an entrepreneur; Peters, 2016). Thus, although there is no purely neoliberal society or state – neoliberalism evolves within various societies in different ways (see Harvey, 2005) – neoliberal political theory allows us to clarify the political premises that underlie the disparate versions of neoliberalism.

In preserving the political state, neoliberal individualistic premises do not accommodate the notion of the people, i.e., the citizens of a given political community or a unitary political body (*demos* or *populus*), understood as an *ultimate intentional lawmaker* or *sovereign* (Locke, 1679 (1960)). The category of the people is a political criterion, which refers to the main act of the people's sovereignty: their giving law to themselves, in the form of rights and duties (Locke, 1679 (1960); Kant, 1793 (1977); Rousseau, 1762 (1964); Sieyes, 1789 (1989)). Putting to the side the relationship between political (Dahl, 1998; Rawls, 1999; Sieyes, 1789 [1989]) and ethnic (Habermas, 2000, 2008) criteria, this act unifies individuals who belong to different ethnicities, cultures, and linguistic traditions. The results of this act are the civic, political and social human rights which have traditionally been the privileged content of the laws of peoples (Locke, 1679 (1960); Kant, 1793 (1977); Marshall, 1950; Rawls, 1971, 1999).

It is true that women and slaves have historically been excluded from the category of the people. It is also undeniable that such exclusion has not been completely overcome and that new categories of exclusion have emerged, such as ageism and digital exclusion. Important political differences within peoples on the axes of class (Badiou, 2016), gender (Elstain, 1981), race (Wilson, 2012), and citizenship remain. Nonetheless, the content of the laws of peoples has provided political criteria for denouncing and reducing, if not eliminating, these exclusions (e.g., in South Africa with the end of Apartheid).

Despite the complexity of the relationship between the state and the sovereignty of the people (Habermas, 2008), the political criterion stresses the subordination of the state to the sovereign people. It also points to the reformulation of the powers of states, "specifying that their legislators must not make certain laws, or must advance certain objectives" (Pyke, 2001, p. 205). For example, instead of exclusively preserving peace or economic and financial efficiency, states ought to ensure the well-being of their citizens. In the absence of such restrictions, the overestimation of states' economic goals (such as low inflation, the removal of trade barriers and foreign currency control, and minimal regulation of the economic labor market) can result in the undermining of welfare at the national (Brodie, 2007) and international level (Beck, 2002).

Some argue that nation states provide a criterion for determining political belonging (Miller, 2000). However, the political criterion points to the fact that one's relation to a given nation state should be based on common laws, not ethnic or cultural differences. Rawls's, (1999) liberal approach to international relationships argues against cosmopolitan principles of justice that are blind to the political (and moral) differences between peoples, for example the difference between liberal and decent peoples, where the former is based on an individualistic tradition and the latter on a 'corporative' tradition. Despite the perils of extending sovereign power to the global order (e.g., populism) and people's incomprehension of the full import of economic and political factors, this order should respect the sovereignty of peoples. Neoliberalism's "global policy of boundary removal" (Beck, 2002, p. 78) undermines the sovereignty of the people (Beck, 2002; Overbeek, 1993). Indeed, the growth of international law affects domestic legal systems, limiting the political choices of legislators and voters, and competition in globalized markets does not allow nations or states to regulate their industries and workplaces. As Hickel notes, for example, financial liberalization creates conditions under which "investors can conduct momentby-moment referendums on decisions made by voters and governments around the world, bestowing their favor on countries that facilitate profit maximization while punishing those that prioritize other concerns, like decent wages" (Hickel, 2016, p. 147).

Peoples are the main 'actors' in the international and global arena, their sovereignty, along with their constitutional power, cannot dispense with common laws. Despite the crucial issue of the existence of mechanisms for enforcing those laws, human rights such as freedom from slavery and serfdom, mass murder and genocide can provide their content (Rawls, 1999). Although the political manipulation of the law by national-hegemonic principles (Beck, 2002) and the enforcement issue (Lane, et al. 2006) must be kept in mind, the human rights approach is relevant to Locke's and Kant's concepts of the people. There is a difference between the national order underlying Locke's and Kant's approaches to the sovereignty of the people and our contemporary international and global order, human rights can create, at the national, international and global level, a sense of political belonging (Habermas, 2008; Lane et al. 2006; Rawls, 1999). As political criteria, human rights preclude resolving persistent political conflicts on the basis of ethnic or national criteria, as occurs with populism and nationalism, respectively.

Given this intricate theoretical framework, as well as the complexity of the notion of a sovereign people (Butler, 2016; Morgan, 1988; Morris, 2000), we stress that whatever its scope, *the sovereign people plays a protective role with regard to citizens' liberties in general and against despotic power in particular* (Locke, 1679 (1960); Kant, 1793 (1977)). Locke, (1679 (1960)) and Kant, (1793 ([1977)) assume that the sovereign people guarantees individual liberty in any human association. Both thinkers hold both that human associations (or societies) of free persons cannot deny the political facts of power, obedience and command (Locke, 1679 ([1960); Kant, 1793 (1977)) and that, in natural (rather than political) conditions, individual liberty is unrestricted. Since in the state of nature it is possible for one to obey unconditionally, having only duties, while the other in turn commands unconditionally, having only rights, the unrestrictedly obedient enjoy no protection against unrestricted power, at least concerning their right to life (Locke, 1679 ([1960); Kant, 1793 (1977)). From this perspective, i.e., from the perspective of individual liberty, the practical (as opposed to theoretical) challenge consists in conceiving of an alliance between individuals that does not undermine their individual liberty. The people as a political body expresses precisely this alliance: an inter-protective construction that replaces the state of unconditional obedience and command.

Following the controversial model of the contractual act (Gough, 1957), individuals transfer to the political power their unrestricted natural right to liberty. This transfer transforms them into "one people, one body politic" (Locke, 1679 (1960), II, p. 89). As members of the people, individuals equally consent to restricting their liberty under a political order and to preserving an equal coercive power, which prevents them from being reduced to servile persons and, correlatively, prevents any one of their numbers from becoming a despotic lord (Locke, 1679 (1960); Kant, 1793 (1977)). As such, they establish *public law* – a system of laws for a people, i.e., an aggregate of human beings, or an aggregate of peoples (Kant, 1793 (1977)) – which allows them to live in a lawful state.

Through public law, i.e., laws based on their will, the people provides to each individual a unique set of liberties with regard to the use of material goods and imposes on each a unique set of restrictions (Locke, 1679 (1960); Kant, 1793 (1977)). When pursuing their personal well-being, as members of the people, individuals cannot ignore this common set of rights and restrictions. When pursuing their well-being, individuals are also, but not exclusively, bound to demands that are independent of their individual interests.

Public vs. private law

Neoliberal theory and practice does not preclude a common law (Buchanan and Tullock, 1962; Hayek, 1976). The common law that it involves is not, however, a law of the people that provides liberties (rights) and imposes a unique set of restrictions (Buchanan and Tullock, 1962; Hayek, 1976; Nozick, 1974). Indeed, neoliberal political theory does not allow for the transformation of individual personalities or isolated natural selves into a collective or single public, viewed as the ultimate intentional lawmaker, which is the model we find, for example, in Locke, (1679 (1960)), Kant, (1793 (1977)), and Rawls, (1971). In Nozick's political theory, when private persons establish a contract to govern their use of the possessions over which they have a private right (Nozick, 1974) – this conception of rights includes both material possessions and natural talents – they are always separate units that remain separate even when they form associations (Nozick, 1974). They do not constitute a common person subject to common legislation that defines and regulates political authority and applies equally to all persons. This mirrors Hayek's suggestion that it is absurd to speak of rights as claims which no one has an obligation to obey, or even to exercise (Hayek, 1976). On this view, human rights result from personal interests, and persons cannot be bound to claims that are independent of their private interests. These claims presuppose a public obligation (or the possibility of coercion), which involves a political organization in which decision-makers act as collective agents: as members of a people rather than individuals. Yet on the neoliberal conception, collective deliberation of this sort limits, and even undermines, individual liberty (Buchanan and Tullock, 1962; Hayek, 1976; Nozick, 1974), leading to oppression (Buchanan and Tullock, 1962), if not to serfdom (Hayek, 1960).

Viewed from the neoliberal standpoint as a *meaningless or mystical* political category (Buchanan and Tullock, 1962) – "a fairy tale" (Hayek, 1960, p. 35) – the political deliberation of the people imposes obligations on individuals, undermining their liberty and well-being. The people as a political body is based on the supposition *that someone* (the people) *can intentionally* prevent or promote certain results, which, via end-rules, guiding organizations can compel individuals to attain. In addition to their "epistemological impossibility" (Gray, 1993, p. 38), however – individuals' multiple interactions produce unpredictable and unforeseen results – end-rules interfere with individual liberty and worsen the positions of all (Hayek, 1976), in particular those who are better off (Nozick, 1974). Interference (or intervention), which is "by definition an [...] act of coercion" (Hayek, 1976, p. 129), is "properly applied to specific orders [that aim] at particular results" (Hayek, 1976, p. 128). Moreover, interference and intervention occurs "if we changed the position of any particular part in a manner which is not in accord with the general principle of its operation" (Hayek, 1976, p. 128).

The general principle of the operation of the spontaneous society is negative liberty, or "the absence of a particular obstacle – coercion by other men" (Hayek, 1960, p. 18) in one's pursuit of maximal individual well-being. Requiring that the situation of the less well off be improved via the principle of the equality of opportunity, for example, involves restricting individual liberty in order to improve the situations of others (Hayek, 1960, 1976; Nozick, 1974). This improvement is thought to be unacceptable because, in addition to presupposing that we can determine the circumstances under which individuals pursue their aims, binding persons to claims that are independent of their private interests constitutes an interference in their liberty (Hayek, 1976). Even if it is admitted that the principle of equal opportunity entails neither complete control over the circumstances in which individuals pursue their well-being (Rawls, 1971), nor equality of results (Rawls, 1971), nor the worsening of the position of the better-off (see Rawls's principle of difference, Rawls, 1971), the fact that it involves *changing the positions of individuals* via a public rule means that it constitutes the imposition of an illegitimate obligation on individuals (Hayek, 1960; 1976; Nozick, 1974). The public law limits the overall sum of well-being – the greater the privatization, the greater the well-being – and restricts the unlimited intensification of individuals' purely private interests (see Hayek's, (1976) and Nozick's, (1974) criticism of the utilitarian and Rawlsian theories of social justice). "Inconsistent" (Hayek, 1976, p. 129) with individual liberties from the perspective of negative liberty and with the unlimited intensification of individuals' purely private interests, public rules are transformed into private rules (commands or end-rules).

On the neoliberal view, the pursuit of individual ends ought to be based on historical principles (Nozick, 1974) or Hayek's abstract rules, which only set out the procedures for acquiring and preserving individual well-being and which do not refer to a common purpose, such as social justice: "Freedom under the law rests on the contention that when we obey laws, in the sense of general abstract rules irrespective of their application to us; we are not subject to another man's will and are therefore free" (Hayek, 1960, p. 11). Under this negative conception of liberty, abstract rules allow for the improvement of "*the chances of all* in the pursuit of their aims"; they are therefore truly *public rules*:

> To regard only the public law as serving general welfare and the private law as protecting only the selfish interests of the individuals would be a complete inversion of the truth: it is an error to believe that only actions, which deliberately aim at common purposes, serve common needs. The fact is rather that what the spontaneous order of society provides for us is more important for everyone, and therefore for the general welfare, than most of the particular services which

> the organization of government can provide, *excepting only the security provided by the enforcement of the rules of just conduct*. (Hayek, 1960, p. 132 emphasis added).

Neoliberal "public" rules are therefore abstract rules that *exclude common concern*. Organizations "sanction" the rights resulting from individuals' interactions under abstract rules (Hayek, 1976). This means not only that governments ought to mirror that order – they cannot provide any rights of themselves – but also that the judicial system ought to be redesigned to fit with the Great Society. Indeed, Hayek critiques the enslavement of law by "false economics" (Hayek, 1960, p. 67), i.e., economics that are dependent on the existence of public goods, and "prophetically" foresees the disappearance of this law in the spontaneous society (Hayek, 1960). Other neoliberal theorists have conceived of the neoliberal impact on law in similar terms, envisaging a legal system based on "true neoliberal economics", which transforms the law into a bond "oblig[ing] one party to behave according to the expectations of the other" (Supiot, 2013, p. 141; see also LeBaron, 2008; McCluskey, 2003; Wacquant, 1999).

This model cannot accommodate the idea of a public person, the people, to whom individuals belong; indeed, the role of ultimate intentional lawmaker is taken from the people and given to the *spontaneous order*, the Great or Open Society. Understood in analogy with the economic market, and equating to abstract rules applied to "an unknown number of future instances" (Hayek, 1976: 35), this spontaneous order constitutes the sovereign lawmaker (Queiroz, 2017).

Neoliberal political intervention under private law

Under the negative conception of liberty, individual freedom is compatible with impediments and constraints (liberty is not bare license, which ultimately undermines negative liberty; Berlin, 1958). Abstract rules allow for private restrictions on liberty, and neoliberal governmental organizations ought to ensure that any restrictions on liberty are limited to the private realm. Neoliberal theorists do not understand this protection as a form of intervention or interference, however. Hayek, (1960), for example, argues for this notion by establishing a distinction between repairing and intervening. When a person oils a clock, they are merely repairing it, securing the conditions required for its proper functioning. In turn, when a person changes "the position of any particular part in a manner which is not in accord with the general principle of its operation" (Hayek, 1976, p. 128), for example by shifting the clock's hands, this counts as intervention or interference. In other words, just as oiling a clock provides the conditions required for its proper functioning, so governmental protection of the private scope of restrictions on liberty allows for the proper functioning of the Great Society. Both merely create the conditions under which individual wellbeing can be maintained, if not increased. In turn, just as shifting the hands of a clock is not in accord with the general principle of the clock's operation, public rules, which impose illegitimate obligations on individuals, constitute an intervention into the functioning of the spontaneous society.

When establishing the particular character of organizations' rules, and excluding "the security provided by the enforcement of the rules of the just conduct" (Hayek, 1960, p. 132), this enforcement means that neoliberal politicians intentionally intervene, but only to prevent the auto-destruction of the "mechanism" itself. They permanently adjust the rules to the neoliberal common law.

Consider a situation in which two people, A and B, are involved in cooperative activity and in which both establish a common rule to safeguard the maximization of their interests. Under this rule, A and B both contribute to the maximization of their own well-being. Although it accepts the interdependence of individuals when pursuing their personal well-being, neoliberal reparation does not allow for a common right to the results of that cooperative interdependence (Hayek, 1976; Nozick, 1974). In denying the existence of a public person, a public will, and in ultimately challenging the idea that there is a common right to a share in the total well-being that results from the contributions of all, neoliberalism not only allows, but also *requires*, that one party has a claim to the exclusively private enjoyment of the benefits of their mutual relationship. Accordingly, neoliberal repair (a metaphor for neoliberal government) ought to remove public law, which allows for the common right to well-being, and should replace it with private law. In this way, the proper functioning of the Great Society – which permits the unrestricted preservation and increasing of individuals' private wellbeing – can be reestablished. The resulting intensification of poverty and inequality (Greer, 2014; Matsaganis and Leventi 2014; Stiglitz, 2013), the diminishing security of employment and income (Clayton and Pontusson, 1998; Stiglitz, 2013), and growing authoritarianism (Brown, 2015; Bruff, 2014; Kreuder-Sonnen and Zangl, 2015; Orphanides, 2014; Schmidt and Thatcher, 2014) are not problems in themselves. To the contrary, to the extent that it undermines individual liberty, any attempt to redress these effects violates the law of the neoliberal state, which, Hayek would say, is based on "true economics". Accordingly, when choosing between the intensification of poverty and inequality and allegiance to the right of non-interference, non-interference must prevail, thus preventing political and social action to reduce (or compensate for) poverty and inequality. Notwithstanding the underlying theoretical debate on the legitimacy and justice of the acquisition of private rights (Hayek, 1976; Marx, 2000; Nozick, 1974; Rawls, 1971, 1993), enforcing the rules of the Open Society *deprives one part of that society of the right to their well-being and to their contribution to the general well-being*. Under the neoliberal model of government and law, certain citizens are deprived of the right to enjoy the public goods that result from their collective activity, while others enjoy a *private right* to goods that result from the contribution of all. Since those who benefit are not able to acknowledge the contribution of others, they erase it and privatize the public law. This privatization shows that the neoliberal trinity of privatization, flexibilization and deregulation ultimately results from the *original privatization of the public or common law*.

Private restrictions on liberty and coercive positive liberty

Aside from the controversy concerning the epistemological value of the distinction between negative and positive liberty (Berlin, 1958 [1997]; Gray, 1993; Rawls, 1971, 1993; Taylor, 1979), theoretical disagreement about their meanings (Taylor, 1979), and the caricatures by which they are often understood (e.g., positive liberty as a form of being "forced-to-be-free"; Taylor, 1979), governmental protection of private restrictions on liberty under neoliberalism shows that neoliberal political theory does not dispense with the coercive feature of positive liberty (see Gray, 1989 for a reading of Hayekian freedom as more than merely negative).

This not a negligible issue; neoliberal political philosophers establish a relationship between the main act of the people's sovereignty, or its constitutional power – establishing a public law that provides to each person a unique set of liberties with regard to the use of material goods and imposes on each a unique set of restrictions – and the violation of individual liberty (Hayek, 1976; Nozick, 1974). The replacement of the people's sovereignty with the spontaneous order is thought to be justifiable because "when we obey laws, in the sense of general abstract rules irrespective of their application to us, we are not subject to another man's will and are therefore free" (Hayek, 1960, p. 11). When arguing against the oppressive nature of the rules that issue from the people, neoliberalism relies on the positive meaning of liberty (freedom to be one's own "master"; Berlin, 1958 (1997)). A *private right to a good* that results from the (perhaps unequal) contribution of all constitutes a coercive act of positive liberty – "coercing others for their own sake, in their, not my, interest" (Berlin, 1958 (1997), p. 397). Similarly, the imposition of that right on society as a whole through legislation, including those who have been deprived of their well-being, also constitutes positive *coercion*. Citizens who are deprived of their well-being must simply accept the *neoliberal diktat*, i.e., the transference of their well-being to the few (Stiglitz, 2013). In a paternalistic way – according to Berlin, (1958 (1997)), positive liberty is always paternalistic in some sense – neoliberal politicians argue that *there is no alternative*

(TINA) to neoliberal political legislation (the government knows best). Consequently, under the veil of state juridical and political violence, neoliberal politicians present governmental rules as an *ultimatum*, precluding consent, i.e., forcing individuals to give up their political right to challenge that deprivation (see the political meaning of *TINA*, Queiroz 2016; Queiroz 2017). The rejection of all public right, i.e., the exclusion of peoples, introduces into the core of the theory (and its practice) the despotic feature that neoliberalism attributes to the general will. In other words, the neoliberal political order mirrors the despotic nature that neoliberals attribute to the *meaningless or mystical* general will (Buchanan and Tullock, 1962).

The neoliberal ultimatum not only protects those citizens who apparently do not need the state's intervention but also ensures that the law only protects their interests (which constitutes the privatization of legal protection). Neoliberal theorists understand public rules as means of protection, as if private interests were not highly dependent on law. Indeed, Nozick's distinction between 'public', "paternalistically regulated" citizens (Nozick, 1974, p. 14) and free citizens, who dispense with state intervention, obscures the existence of private, "paternalistically regulated" citizens. These citizens are protected by the reparations of neoliberal "public" law. In addition, however, rather than accepting the collective protective scope of the law, they demand a monopoly on it. Although neoliberalism casts them as utterly independent actors – lone Robinson Crusoes – they are highly dependent not only on the contributions of others for their well-being but also on the positive law. This shows that unless there is a *common law* to prevent others from interfering with one's liberty and to provide certain means, negative liberty is an empty claim.

Insofar as the protective function of the government and the positive law include both legislative and coercive power, instead of coercing others for one's own sake, neoliberal positive liberty allows private individuals to impose, without consent, public restrictions for the sake of their private interests. Neoliberal positive liberty thus leads to the establishment of legal and political inequality: some command without consent, i.e., without restriction, while others obey without consent, i.e., without liberty. Ultimately, making use of the benefits of negative liberty depends on the (political) attribution to individuals of certain legal and political statuses, under which they can make use of their liberty.

Moreover, the positive liberty that underlies the spontaneous order not only deprives certain citizens of their share of the general well-being but also leaves no room to claim a right against that deprivation. Besides protecting negative liberty in the maximization of individuals' well-being, this order does not provide any concrete rights. Hayek explicitly says that it "is meaningless to speak of a right in the sense of a claim on the spontaneous order" (Hayek, 1960, p. 102, II). Indeed, although framed by abstract rules, rights are always obtained under particular circumstances, i.e., in terms of differences between "individuals", for example natural and social talents (Hayek, 1976; Nozick, 1974). Despite the interdependence of all individuals, individuals always remain separate unities and are thus deprived of the right to claim a common share of the fruits of their relationships – as if belonging to a common body entailed personal indifference and the abandonment of private interests. Accordingly, if the Great Society, which replaces the will of the people, does not provide rights to citizens, and if those citizens do not obtain them from their private interactions, it is meaningless to claim such a right or to complain that such a right has been denied them. *There is nothing to claim or to complain about*. In other words, where there are no rights, there can be no deprivation of rights.

Even if individuals wish to complain about the deprivation of their rights, the neoliberal state – which considers such rights imaginary, fictitious, mystical – does not contain institutions that can address such complaints. Under the neoliberal state, both the people and public institutions vanish into thin air. As Beck stresses with regard to neoliberal globalization, neoliberalism is the power of Nobody (Beck 2002). Alluding to Odysseus's clever escape from the cyclops Polyphemus in the Odyssey (Homer, 1996, 9, pp. 414–455), Beck suggests that the Nobody created under neoliberalism does not establish, protect or enforce equal individual rights. Even though Nozick (unlike Hayek) accepts the existence of natural rights and liberties, his rejection of a public person and public restrictions shows that the assumption of natural rights does not guarantee their enjoyment. In other words, when the will of the people becomes a mirage, individuals' natural rights are also rendered illusory, as the neoliberal spontaneous society illustrates. Accordingly, instead of allowing for the "creat(ion of) conditions likely to improve the chances of all in the pursuit of their aims" (Hayek, 1976, p. 2), private restrictions on liberty deprive certain citizens of the chance to pursue their aims (Brown, 2015; Gill, 1998; Hall, 2011; Klein, 2007; Overbeek, 1993; Stiglitz, 2013, 2016). Instead of protecting individual liberty, the rejection of the "fairy tale" of the people allows for the emergence of two familiar political statuses, originally deployed in neoliberal political society: those who live under free self-serfdom on the one hand and the invisible and voiceless on the other.

Free self-serfdom and voiceless persons

A free serf is someone who, although deprived of political protection – whether this is understood as it was in the medieval era (Bloch, 1961), which made a distinction between the protector and the protected, or as it was understood in the liberal tradition (Locke, 1679 (1960); Kant, 1793 (1977)), in which each person is simultaneously protector and protected – can still satisfy their bodily needs through selling themselves or their labor. Neoliberal private restrictions on liberty cannot override the unrestricted autocratic deliberation of those who, in the absence of public law, can freely renounce their liberty in situations of extreme need, thus voluntarily enslaving themselves. The rejection of a public limit to individual liberty, along with the overlapping of public law and private interests, allows for unrestricted orders and, correlatively, for obedience without liberty (on work precariousness see Gill and Pratt, 2008; on work conditions in sweat shops, see Bales 1999). Consequently, neoliberal political theory and practice allow for the creation of a situation in which some citizens (serfs) only obey while others (lords) only command.

One may argue that despite social and economic differences, along with their non-negligible impact on individual liberty (Marx 2000; Rawls, 1971), neoliberalism's Great or Open Society is not compatible with serfdom. Regardless of the lack of clear political criteria for defining an individual's legal and political status (Bloch, 1961), human relationships have evolved under conditions of legal and political inequality (for example the superior free person vs. the inferior serf or vassal). This legal and political inequality is at work, for example, in systems where lords offer protection in exchange for total obedience (on the part of serfs and vassals) (Bloch, 1961). From the perspective of neoliberal theory, we are all equal: neoliberal society does not contain legal or political inequality and does not divide citizens into those who are superior and those who are inferior. It also does not include "protective relationships" or juridical and political obligations. To be at the disposal of someone else who can do whatever they please and to whom one owes unrestricted obedience entails neither that one has an inferior legal status nor that the political relationship at stake is one of a superior to an inferior. Persons have the same legal constitutional status (they all are seen as equally free), and all are equally entitled to pursue their private interests. Even if people sell themselves, this concerns the private restriction of liberty from the perspective of neoliberalism and does not conflict with the conditions required for the proper functioning of the spontaneous order, i.e., with individuals' private liberty. Still, the private scope of individuals' mutual service – the forbidding of *serving others for the sake of those others' well-being* – does not prevent a person's serving *another* as a means of ensuring their own private wellbeing, in which case it would not be appropriate to understand their relationship in terms of servant and seignior.

Besides entailing what is known in political philosophy as the liberty of slaves, i.e., the liberty of choosing either to comply with the orders of the master or to be beaten to death, the privatization of the well-being that results from individuals' cooperation is based on the coercive restriction of liberty, under which some obey without liberty and others command without restriction.

Thus, even if in neoliberal spontaneous societies people are not assigned explicitly different political statuses, which entail different political rights and duties, neoliberal political society does not prevent people from becoming servile or, correlatively, from becoming despotic. This fact reveals the extent to which neoliberalism entails a dangerous process of what some authors have called refeudalization (Supiot, 2013; Szalai, 2017), full analysis of which deserves examination of its own.

Nevertheless, when *obeying without liberty*, if citizens fail to acquire their rights they risk becoming something *less than* a free serf, i.e., a free excluded citizen. A free excluded citizen is a citizen who lives in a free society *without having the personal, social or institutional resources to make use of their own liberty*. When the neoliberal spontaneous order does not provide any concrete rights, and when another's wellbeing has no bearing on one's own, one is unrestrictedly free to pursue one's own wellbeing even to the detriment of others unilaterally (the fully alienated person can be thrown away). In this case, voiceless and invisible citizens can only enjoy purely negative liberty, in the absence of the personal, social and institutional resources with which they might otherwise achieve well-being. Neoliberalism also entails the continuous risk of passing from servile (or docile) citizenship into lawless personhood. As such, individuals' social existence is excluded from the neoliberal subjectivation procedure itself (in which human beings make themselves and are made subjects, Foucault, 2008).

Neoliberalism does not reduce to fostering the entrenchment of political inequality: the division of citizens into those who obey and those who command. It also does not merely imply a situation in which some are protected by the state while others are not, where private interests have a monopoly on legal protection and rights while others are denied political protection and only have duties (on work precariousness see Gill and Pratt, 2008). Similarly, it does not exclusively entail political arbitrariness; the private reduction of the "public" law allows for the unilateral institution of the rules (or their revocation). Ultimately, neoliberalism risks leading to the total exclusion of some citizens *under the veil of full liberty*. The vanishing of the will of the people results in the invisibility of certain kinds of people, who are then forced to live in the spontaneous society as if they were stateless or lawless persons.

It is true that under the distinction between neoliberal theoretical premises and neo-liberal practice individuals' lack of protection does not correspond to these extreme cases. There is a distinction between neoliberal theoretical premises and neoliberal governmental laws within the many versions of the welfare state, for example neoliberalism's reshaping of previous (welfare) state policies along neoliberal lines (Kus, 2006). Neoliberalism has retained some of the elements of that state (such as the protection of the rights of the most vulnerable), although these elements have been reshaped by the market approach to social welfare

(Hartman, 2005; MacLeavy, 2016). On this basis, neoliberal officials have assigned public goods and services to private market providers, redesigning social programs to address the needs of neoliberal labor markets rather than personal wellbeing and establishing partnerships between the state and the private sector (Brodie, 2007).

Moreover, some argue that neoliberalism's market approach to social welfare was an attempt to overcome certain economic and social difficulties of the welfare state. For example, economic internationalization has affected the competitive viability of the welfare state (Boyer and Drache, 1996; Rhodes, 1996). Also, the expansion of the state weakened intermediate groups and jeopardized individual liberties, subjecting citizens to increasing bureaucratic controls (Alber, 1988). We shall not dwell on a full analysis of these developments. The neoliberal market approach is, however, incompatible with the very idea of a welfare state. Indeed, despite the differences between the socialist, conservative and liberal versions of that state (EspingAndersen, 1990), welfare states protect social rights, such as the right to education and health, and therefore provide social policies to enforce them (Marshall, 1950; Esping-Andersen, 1990), such that "[t]he provided service, not the purchased service, becomes the norm of the social welfare" (Marshall, 1950, p. 309). Moreover, the functioning of the welfare state requires the contribution of fellow citizens (Marshall, 1950; Esping-Andersen, 1990). By contrast, the market approach rejects in principle all social rights, such as the right to education and health, and requires that individual welfare be an exclusively private enterprise (Brodie, 2007; MacLeavy, 2016). Instead of being provided, such services ought to be purchased (Brodie, 2007; MacLeavy, 2016).

Moreover, if the economic market only identifies solvable needs, and if individuals cannot signal their lack of resources, the neoliberal welfare state cannot prevent individuals who have been deprived of their rights from becoming invisible, along with the resulting institutionalized insecurity (Brodie, 2007), intensified poverty and inequality, and diminishing of security of employment and income for many wage earners (Clayton and Pontusson, 1998; Stiglitz, 2013). If the spontaneous society and its governments do not provide any rights, and if individuals do not acquire them in the economic market, there is no reason to claim such rights (including social rights). In this case, neoliberal social welfare reduces to charity (Clayton and Pontusson, 1998; Raddon, 2008; Mendes, 2003). Under this reduction, neoliberal theory fosters individuals' dependence on the private goodwill of citizens who, after legislating with their own interests in mind, and after denying others the *right* to enjoy the fruits of their own contributions, then establish government spending as a "free lunch" of sorts (all the while paradoxically arguing that "government spending is no free lunch" (Barro, 2009); see Nozick's, (1974) defense of charity)). The neoliberal conception of welfare also shows how neoliberal theory and practice do not prevent the subordination of certain individuals to nonconsensual external mastery.

Neoliberalism is equally committed to state retrenchment or permanent austerity (Whiteside, 2016). By requiring fiscal consolidation, cuts to social security, the privatization of public property, the liberalization of collective bargaining, and the shrinking of pensions (Barro, 2009), austerity not only undermines all attempts to establish social security but also challenges the liberal and democratic basis of society. First, neoliberal austerity neglects people's well-being. A Portuguese neo-liberal politician declared in 2013 that even if under austerity measures the well-being of the people had worsened, the country was better off.[1] The fact that neo-liberal policies have improved the state market is more relevant than the fact that the Portuguese people have been neglected and severely harmed (Legido-Quigley et al. 2016).

Second, neoliberalism excludes in principle the will of the people, i.e., it obliges citizens to obey private laws to which they have not consented. Consequently, it excludes citizens' rejection of its harmful effects, such as poverty and inequality, and rejects all appeals to alternative policies. Following the political referendum of 2015, for example, where the people voted against neoliberal politics of austerity[2], the Greek government nonetheless imposed a third harsh and austere economic program[3].

[1] http://www.jn.pt/live/entrevistas/interior/a-vida-das-pessoas-nao-esta-melhor-mas-opais-esta-muito-melhor-3697968.html

[2] https://www.theguardian.com/business/2015/jul/05/greece-referendum-early-resultshistoric-no-vote-against-austerity, Accessed 14 Feb.

[3] https://www.mintpressnews.com/before-the-brexit-greek-voters-said-no-to-austeritymeasures-got-more-austerity-measures/218122/; Accessed 14 Feb

Accordingly, neoliberal political principles, embedded in austerity policies, cannot prevent certain citizens from becoming *invisible* and voiceless citizens, i.e., *Nobodies*. As voiceless citizens, their preferences can only be registered through illiberal and antidemocratic channels, such as populism. Only following the election of US President Trump did the deteriorating life conditions of American citizens living in the rust belt states of Michigan, Pennsylvania and Wisconsin become widely known (Walley, 2017). Treated as nothing, and having becoming Nobodies, these citizens face the oppressive and violent institutional neoliberal Nobody, with its no less violent and oppressive political body.

The rise of populism

There is a lack of consensus on the definition of populism (Collier, 2001). It can, however, be described as an organizational or a strategic approach (Weyland, 2001) and ideology (Freeden, 2016; MacRae, 1969; Mudde, 2013; Mudde and Kaltwasser, 2013). The organizational perspective of populism stresses the importance of the personal leader, who bases his or her power on direct, unmediated, and institutionalized relationships with unorganized followers (Weyland, 2001). In turn, as an ideology, i.e., a set of beliefs, values, attitudes, and ideas, populism combines (not always coherently and clearly) political, economic, social, moral, and cultural features with several characteristics that appear together, such as emphasis on the leader's charisma: "the populist can demand the highest principles in the behavior, moral and political, of others while being absolved him or herself from such standards" (MacRae, 1969, p. 158). Beyond these features, however, and despite the fact that the concept of the "pure" people and the corrupted elite can be framed in different ways (Canovan, 1999), the *pure* and homogenous *people* and the *corrupt* and homogenous *elites* are core concepts that underlie populist ideology (Mudde, 2004).

Since neo-liberal officials do not consider citizens' and peoples' political claims and are not entitled to address the political, economic, and social consequences of their policies, the perception that neo-liberal politicians are corrupt elites has been on the increase (Mudde and

Kaltwasser, 2013). This has helped populist leaders to replace neo-liberal politicians, allowing populism to fill the emptiness that has resulted from the failure of those in power to address the people's claims.

Although the relationship between neoliberalism and populism deserves its own examination, the exclusion of the people, along with the right to reciprocal coercion, is a point of tacit agreement between neoliberalism and anti-liberal, anti-democratic political forces (Weyland, 1999). Populist leaders have employed modern, rational models of economic liberalism – such as fiscal consolidation, cuts to social security, the privatization of public property, the liberalization of collective bargaining, and the shrinking of pensions to undermine intermediary associations, entrenched bureaucrats and rival politicians who seek to restrict their personal latitude, to attack influential interest groups, politicians, and bureaucrats, and to combat the serious crises in Latin America and Eastern Europe in the 1980s (Weyland, 1999). In turn, neoliberal experts use populist attacks on special interests to combat state interventionism and view the rise of new political forces, including populists, as crucial for determined market reform (Weyland, 1999). We therefore ought to be careful not to criticize neoliberal authoritarianism while neglecting the hidden powers that secretly support neoliberalism's disdain for the people, such as mafias (Schneider and Schneider, 2007). Indeed, those who do so may take pleasure in seeing the blame for authoritarianism fall exclusively on the shoulders of neoliberal theory and practice, even though they too endorse a form of governance and the administration of the state apparatus that does away with the people.

When individuals' relationships evolve in the absence of the people and of laws to protect against despotic and abusive power, an increase in illiberal and antidemocratic forms of resistance to neoliberal policies can only be expected (Gill, 1995; Hickel, 2016). As Locke, (1679 (1960): II, p. 225) put clearly:

> Great mistakes in the ruling part, many wrong and inconvenient Laws, and all the *slips* of human frailty will be born by the People, without mutiny or murmur. But if a long train of Abuses, Prevarications, and Artifices, all tending the same way, make the design visible to the People, and

> they cannot but feel, what they lie under, and see, whither they are going; 'tis not to be wonder'd, that they should then rouze themselves, and endeavour to put the rule into such hands, which may secure to them the ends for which Government was at first erected.

If we accept that (a) impoverishment and inequality are on the increase; (b) governments are refusing to provide political remedies for this impoverishment; (c) and citizens' political choices are being neglected in a long series of abuses, it is not surprising that voiceless citizens may try to put the ruling power into illiberal hands that will achieve the purpose for which government was first established: securing the common public good. Under the neoliberal transformation of private rules into public rules, citizens are witnessing a continuous disregard for their collective well-being (see the relationship between the election of Donald Trump and the deteriorating life conditions of American citizens living in the rust belt states of Michigan, Pennsylvania and Wisconsin; Walley, 2017).

Instead of welcoming populist reactions, however, we should be clear that the anti-liberal and antidemocratic hijacking of the citizens' revolt against neoliberalism in no way respects the need for public rules. A call for the establishment and protection of public law is a call for *personal and institutional liberal and democratic sovereignty*, which differs fundamentally from populism and the neoliberal model of sovereignty (Dean, 2015; Foucault, 2008). This claim also rejects the political (and nightmarish) choice between neoliberalism and populism. Indeed, even if the relationship between liberal democracy and populism deserves investigation of its own, *liberal and democratic sovereignty* does away with the distinction between the pure and homogenous people against corrupt and homogenous elites. It also rejects the idea of the personal and benevolent leader/ protector, who bases their power on direct, unmediated, and institutionalized relationships with unorganized followers.

First, although the distinction between corrupt elites and the pure people rightly points to the problem of the legitimacy of the rulers' power, the people is not a homogeneous or pure body, whatever the criterion of belonging (ethical, ethnic, racial, economic). Far from referring to an undifferentiated and homogeneous *corpus*, the people is a heterogeneous political body, which includes gender, racial, and economic differences (along with disagreement about personal and collective ends), and which ultimately entails non-alienable individual rights and duties (Locke, 1679 (1960); Kant, 1793 (1977); Sieyes, 1789 (1989)).

Second, the solution to this gap is not its elimination through the immediate relationship between the leader and the pure, homogeneous people. In the liberal political tradition, there is no immediate political power. Rawls's, (1993) political liberalism, for example, points to the gap between the political principles of society (e.g., the principles of justice), which are embedded in its basic political institutions (e.g., constitutions) and in "executive" institutions (parliaments, courts, governments), and the individuals in everyday life. Accordingly, the sovereignty of the people ultimately means that, whether at the political, local, national, international, or global level, citizens' relationships are always mediated by law embedded in their public institutions (Locke, 1679 (1960); Kant, 1793 (1977); Rawls, 1993).

Even if there are many points of ideological disagreement concerning the concept of the people, sparked mainly by its use by controversial figures from the standpoint of liberalism, such as Rousseau's concept of the general will, in Locke's and Kant's political philosophy the sovereignty of the people does not mean that the people can pursue its immediate and unbridled wishes. A charter of rights or constitutional principles always binds the will of the people (Locke, 1679 (1960); Kant, 1793 (1977)). In the absence of such restrictions, the people can itself become a despot, a danger which has been acknowledged since at least the time of Aristotle, (2002; see also Cicero 1999; Locke, 1679 (1960); Rawls, 1971, 1993).

Third, in Locke's and Kant's political philosophies, the protective role of the people aims to ensure a political society of free and equal persons, not a society of minor and inferior subjects who need benevolent protectors, such as populist leaders (see Locke's claim concerning the constitutional protection of individuals' political rights (Locke, 1679 (1960)) and Kant's rejection of paternalistic and despotic political power (1793 (1977)).

Liberal theory challenges the underlying neoliberal and populist Manichean opposition between personal interests and the general will of the people ("either there is a general will or individual liberty is repressed", "if there is individual liberty, the general will is excluded"). If, when protecting the homogenous people against corrupt elites, populists endorse the first alternative, and if the neoliberal exclusion of the people corresponds to the second, both approaches remain blind to the political responsibility of free persons. Ultimately, whether by imposing on others the unrestrictedly and selfish pursuit of wellbeing or by appealing to the unlimited will of the people, both undermine individuals' political freedom.

For these reasons, personal and institutional liberal and democratic sovereignty is more than a childish claim to state protection against political irresponsibility and blindness to public contributions to individual private well-being. It is a claim to one's own political responsibility, for oneself and others, as this claim is clearly formulated in Locke's and Kant's political philosophies.

The social safety net

Although Locke's and Kant's political philosophies do not require individuals under public law to positively foster others' social, economic and cultural well-being, their perspectives on the public challenge indifference towards the increasing poverty and inequality that we are currently witnessing under neoliberalism (Greer, 2014; Stiglitz, 2013). They also speak against the state authoritarianism that neoliberalism engenders (Brown, 2015; Bruff, 2014; Kreuder–Sonnen and Zangl, 2015; Orphanides, 2014; Schmidt and Thatcher, 2014). Of course, we may disagree on the extent of the success or failure of Locke's and Kant's theoretical political constructions of a political personality, understood in analogy with a single body. Some criticize the illiberal nature of Kant's general will (for example the representatives' betrayal of the people's interest in the liberal social contract; Badiou, 2016). Nevertheless, these weaknesses challenge neither individual liberty, nor the people, nor the inter-protective role of the people and public law. Indeed, they remind us of the political meaning of 'the body politic'.

Despite their strong commitment to the protective role of the people, along with their *awareness of our political responsibility for the fairness of the public rules that affect us all*, Locke and Kant do not fully explain the necessity of the notion of the people when it comes to producing a social safety net created by the will of the sovereign people. They also do not consider democratic procedures for arriving at collective support for a social safety net. With the differences between ancient and modern democracies acknowledged (Bobbio, 1988), the fact that Locke and Kant endorse democracy's core feature, the existence of a people (the entire body of citizens) with a right to make collective decisions (Bobbio, 1988), does not make them democrats, at least in our modern sense (Bobbio, 1988).

Following our premises, and acknowledging the various ways in which globalization impacts states and people, democratic governments should establish democratic procedures at the national and international level to secure collective support for the political and social safety net. These include public laws based on the will of the people that provide each person with a unique set of liberties with regard to the use of material goods which impose on each a unique set of restrictions. These liberties and restrictions will ensure that individuals have an *equal coercive power* to prevent their becoming servile persons and, correlatively, to prevent any one of them from becoming a despotic lord. They also require the assumption of the cooperative nature of individual well-being, and therefore the pursuit of social justice with regards to the fruits of that cooperation. The political translation of the *common right* to the results of social cooperation through public policies that protect social rights, such as the right to education and health, is also desirable. This requires the "direct or indirect participation of citizens, and the greatest possible number of citizens, in the formation of laws" (Bobbio, 1988, p. 38). Again, it is necessary to recast the political principle of provided (not purchased) services as a norm of public and social welfare. Finally, it requires awareness of the fact that in the absence of a political body to protect and enforce individual liberties, individuals will lack the *personal, social and institutional resources to make use of their own liberty*.

Conclusion

We have shown that neoliberalism's rejection of the existence of the people seriously harms individual private liberty and does not prevent the transformation of the majority of free individuals into servile persons. More specifically, we have shown that forbidding the public restriction of liberty (which is inherent in the concept of the people) while exclusively defending private restrictions of liberty (a) deprives the majority of citizens of the equal right of coercion, and therefore of equal liberty, and (b) promotes the rise of different political statuses, a division between those who obey and those who command. We have also shown that neoliberalism lacks the resources to prevent the total alienation of liberty.

In comparing neoliberalism to Locke and Kant's political philosophies, we have shown how the protective role of the people is compatible with individual liberty. Since it requires an equal right of coercion, it allows for the protection of individual liberty. We have also shown that this is not an exclusively collective task. It also depends on each citizen. In Locke's and Kant's political philosophies, the protective role of the people aims to ensure that political society is free and equal, not a society of minor and inferior subjects who need benevolent protectors (Locke, 1679 (1960)); Kant, (1793 (1977)). We concluded that, against neoliberalism's faith in the powers of the spontaneous order, individual political autonomy depends on the public safeguarding of liberties. We also pointed out that unless there is a political turn toward the acknowledgement of the people or peoples, along with recognition of the significance of their political deliberation, neo-liberalism cannot be separated from illiberal and antidemocratic political choices. Similarly, if individuals' relationships evolve beyond the existence of the people and lack laws to protect against despotic and abusive power, we cannot prevent the development of slavish and servile relationships among citizens. The fact that these relationships remain politically forbidden in neoliberal states, for example in the European Union, only reveals that neoliberalism's dismantling of liberal and democratic political institutions has not fully succeeded. In the absence of the people, human rights depend exclusively on individuals' interests; the spontaneous order thus cannot prevent neoliberalism from descending into slavery and serfdom, i.e., self-slavery and self-serfdom.

Future research should ascertain how, in the aftermath of neoliberalism's devastating social and political effects on public cohesion, it might be possible to reconstitute a sense of political belonging (Habermas, 2008) and the sovereignty of the people (Pyke, 2001) under globalization.

Future research should also continue to evaluate the dangerous process of what many are calling refeudalization under neoliberalism (Supiot, 2013; Szalai, 2017). It is worth comparing the feudal alienation of political liberty, for example the different perspectives on vassalage (Bloch, 1961), with contemporary forms of inferior political status.

Finally, future research could evaluate how, as a reaction to the disenchantment with the rise of bureaucracy identified by Weber, (1978), neoliberalism might express a kind of re-enchantment with the exclusively individual rational actor, who claims a non-alienable space of liberty against the bureaucratic "iron cage".

References

Alber J (1988) There a Crisis of the Welfare State? Cross-National Evidence from Europe, North America, and Japan. *Eur Sociol Rev 4*(3):181–207

Aristotle (2002) *The Politics* (trans. Carnes Lord). University of Chicago Press, Chicago

Badiou A (2016) Twenty-Four Notes on the Uses of the Word "People". In: Allen A (ed) *What is a People*. (New directions in Critical Theory). Columbia University Press, New York, pp 21–31

Barro R (2009) Government spending is no free lunch: how Democrats are peddling voodoo economics. *Wall Street Journal*. http://www.wsj.com/articles/ SB123258618204604599 Accessed Nov 2015

Bales, K (1999) Disposable People: New Slavery in the Global Economy. University of California Press:Berkeley and Los Angeles, California and London, England Beck U (2002) *Power in the global age*. Polity, Cambridge, Malden

Beetham D (1992) Liberal democracy and the limits of democratization. *Polit Stud Special Issue* 4XL:40–53

Berlin I ((1958) [1997]) Two concepts of liberty. In: Goodin R, Pettit P (eds) *Contemporary political philosophy*. Blackwell Publishers, Oxford, pp 391–417

Bloch M (1961) *Feudal Society: social class and political organization*. The University of Chicago Press, Routledge & Kegan Paul, London

Bobbio N (1988) *Liberalism and democracy*. Verso, London, New York

Boyer R, Drache D (eds) (1996) *States against markets: the limits of globalization*. Routledge, London

Brodie J (2007) Reforming social justice in neoliberal times *Stud Social Justice 1* (2):93

Brown W (2015) *Undoing the demos: neoliberalism's stealth revolution*. Zone Books, Cambridge

Bruff I (2014) The rise of authoritarian neoliberalism *Rethink Marx 26*(1):113–129

Buchanan J, Tullock G (1962) *The calculus of consent: the logical foundations of constitutional democracy*. Liberty Fund, Indianapolis

Butler J (2016) "We, the People" Thoughts on Freedom of Assembly. In: Allen A (ed) *What is a People*. (New directions in critical theory). Columbia University Press, New York, pp 49–64

Canovan M (1999) Trust the people! populism and the two faces of democracy *Polit Stud 47*:2–16

Cicero (1999) *On the commonwealth and on the laws*. Cambridge University Press, Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Clayton R Pontusson J (1998) Welfare-State Retrenchment Revisited: Entitlement Cuts, Public Sector Restructuring, and Inegalitarian Trends in Advanced Capitalist Societies and Source World *Polit 51*(1):67–98

Collier R (2001) Populism. In: Smesler NJ, Baltes PB (eds) *International encyclopaedia of social and behavioral sciences*. Elsevier, Oxford, pp 1813–16

Dahl R (1998) *On Democracy. With a New Preface and Two New Chapters by Ian Shapiro*. Yale University Press, New Haven & London

Dean J (2008) Communicative Capitalism: Circulation and Foreclosure of Politics. In: Boler M (ed) *Digital Media and Democracy: Tactics in Hard Times*. The MIT Press, Cambridge, Massachusetts. London, England, pp 101–121

Dean J (2015) Neoliberalism's Defeat of Democracy. *Critical Inquiry*. http://criticalinquiry.uchicago.edu/neoliberalisms_defeat_of_democracy

Elstain JB (1981) *The public man, private woman: woman in social and political thought*. Princeton University Press, Princeton, NJ

England K, Ward K (2016) Theorizing neoliberalization. In Springe S, Birch K, MacLeavy J (eds) *The Handbook of Neoliberalism*. Routledge, London, New York, pp 27–38

Esping-Andersen G (1990) *The three worlds of welfare capitalism*. Polity Press, Cambridge

Foucault M (2008) *The birth of biopolitics*. Palgrave Macmillan, New York

Freeden M (2016) After the Brexit referendum: revisiting populism as an ideology *J Political Ideol 22*(1):1–11

Gill R, Pratt A (2008) Precarity and cultural work in the social factory? immaterial labour, precariousness and cultural work *Theory Cult Soc 25*(7–8):1–30

Gill S (1995) Globalisation, market civilisation, and disciplinary neoliberalism Millenium: *J Int Stud 24*(3):399–423

Gill S (1998) European governance and new constitutionalism: economic and Monetary Union and alternatives to disciplinary Neoliberalism in Europe *New Political Econ 3*(1):5–26

Gough JW (1957) *The social contract: a critical study of its development*. Oxford University Press, Oxford

Gray J (1989) *Essays in political philosophy*. Routledge, London

Gray J (1993) *Post-liberalism. studies in political thought*. Routledge, London

Greer S (2014) Structural adjustment comes to Europe: lesson for Eurozone from the conditionality debates *Glob Social Policy 14*(1):51–71

Habermas J (2000) Crossing globalization's valley of tears *New Perspect Q 17* (4):51–7

Habermas J (2008) Citizenship and National Identity. Global Justice: Seminal Essays. Vol. I. In: Pogge T, Moellendorf D (eds) *Global Responsibilities*. Parangon House St. Paul, St. Paul, pp 285–309

Hall S (2011) The neoliberal revolution *J Cult Stud 25*(6):705–728

Hartman Y (2005) In bed with the enemy: some ideas on the connections between neoliberalism and the welfare state *Curr Sociol 53*(1):57–73

Harvey D (2005) *A brief history of neoliberalism*. Oxford University Press, Oxford, New York

Hayek F (1960) *The constitution of liberty.* Chicago Press, Chicago and London

Hayek J (1976) *Law, legislation and liberty. the mirage of social justice, Vol. II.* The University of Chicago Press, Chicago, London

Hickel J (2016) Neoliberalism and the end of democracy. In: Springer S, Birch K, MacLeavy J (eds) 2016. *The Handbook of Neoliberalism.* Routledge, London, New York, pp 142–152

Homer (1996) *The Odyssey* (trans. by Robert Flages). Viking, New York

Kant I (1793) [1977]) Über den Gemeinspruch: Das mag in der Theorie richtig sein, taugt aber nicht für die Praxis. In: Weischedel W (ed) *Werke in Zwölf Bänden Band XI 1977.* Suhrkamp Verlag, Frankfurt am Main, pp 125–172

Kersting W (1992) Politics, freedom, and order: Kant's political philosophy. In: Guyer P (ed) *The Cambridge Companion to Kant.* Cambridge University Press, New York, pp 341–368

Klein N (2007) *The shock doctrine: the rise of disaster capitalism.* Allen Lane, London

Kreuder-Sonnen C,Zangl B (2015) Which post-Westphalia? International organizations between constitutionalism and authoritarianism *Eur J Int Relat 31* (3):568–594

Krugman P (2012) Europe's austerity madness. *The New York Times.* http://www. nytimes.com/2012/09/28/opinion/krugman-europe-austerity-madness.html Accessed Mar 2017.

Lane J-E (2006) Law and Politics: Reflections upon the concept of a spontaneous order and the EU. In: Kurild-Klitgaar, Koppl P, Birner J (eds) *Advances in Austrian Economics, Vol. 8. The Dynamics of Intervention: Regulation and Redistribution in the Mixed Economy.* Elsevier, Amsterdam, pp 429–440

Legido-Quigley H, Karanikolos M, Hernandez-Plaza S, Freitas C, Bernardo L, Padilla B, SáMachado R, Diaz-Ordaz Ka Stuckler D, McKee M (2016) Effects of the financial crisis and Troika austerity measures on health and health care access in Portugal *Health Policy 120*(7):833–839

Kus B (2006) Neoliberalism, Institutional Change and the Welfare State: The Case of Britain and France *Int Comparative Sociol 47*(6):488–525

LeBaron G (2008) Captive Labour and free market; prisons and production in USA *Capital Class 32*(2):59–81

Locke J (1679 [1960]) Two Treatises of Government. In: Laslett P (ed) *Cambridge texts in the history of political thought.* Cambridge University Press, Glasgow

Lorenz C (2012) If You're So Smart, Why Are You under Surveillance? Universities, Neoliberalism, and New Public Management *Crit Inq 38*(3):599–629

MacLeavy J (2016) Neoliberalism and Welfare. In: Springer S, Birch K, MacLeavy J (eds) *The handbook of neoliberalism.* Routledge, London, New York, pp 252–261

MacRae D (1969) Populism as an ideology. In: Ionescu G, Ernst GR (eds) *Populism: its meaning and national characteristics*. Macmillan Company, London, pp 153–165

Marshall T (1950) Citizenship and social class. In: Goodin R, Pettit P (eds) *Contemporary political philosophy*. Blackwell Publishers, Oxford

Marx K (2000) *Karl Marx: selected writings*. In: David McLellan (ed) Oxford University Press, Oxford, New York

McCluskey M (2003) Efficiency and social citizenship: challenging the neoliberal attack on the welfare state *Indiana Law J 78*(2):783–876

Mendes P (2003) Australian neoliberal think tanks and the backlash against the welfare state. *J Australian Polit Econ 51*:29–56.

Miller D (2000) *Citizenship and National Identity*. Polity, Cambridge

Morgan E (1988) *Inventing the People*. WWW Norton

Morris C (2000) The very idea of popular sovereignty: “we the people” reconsidered *Soc Phil Policy Foundation 17*(1):1–26

Mudde C (2004) The Populist Zeitgeist *Gov Oppos 39*(4):541–63

Mudde (2013) Populism. In: Freeden M Sargent L T Stears M (eds) *The Oxford Handbook of Political Ideologies*. Oxford University Press, Oxford, pp 493–512

Mudde C Kaltwasser C (2013) Exclusionary vs. inclusionary populism: comparing contemporary Europe and Latin America *Gov Oppos 48*(2):147–174

Mudge S (2008) The state of the art. What is neoliberalism? *Socio-Econ Rev 6*:703–731

Nozick R (1974) *Anarchy, State, and Utopia*. Blackwell, Oxford

Orphanides A (2014) The Euro Area Crisis: Politics over Economics *Atl Econ J 42*:243–263

Overbeek A (eds) (1993) *Restructuring hegemony in the global political economy. The rise of transnational nep-liberalism in the 1980s*. Routledge, London, New York

Peters M (2016) Education, neoliberalism, and human capital: homo economicus “as entrepreneur of himself”. In: Springer S, Birch K, MacLeavy J (eds) *The Handbook of Neoliberalism*. Routledge, London, New York, pp 297–307

Pyke J (2001) Globalization: The Bane of Popular Sovereignty. In: Charles J. Round T (eds) *Beyond the Republic; Meeting the Global Challenges to Constitutionalism* 2001. Federation Press, Leichhardt, N. S. W., pp 205-2014.

Queiroz R (2016) Neo-liberal TINA: an ideological and political subversion of liberalism *Crit Policy Stud 10*(4):1–20

Queiroz R (2017) From the exclusion of the people in neoliberalism to publicity without a public *Pal Commun 34*(2):1–11

Raddon M-B (2008) Neoliberal legacies: planned giving and the new philantropy *Stud Political Econ 81*:27–48

Ramey J (2015) Neoliberalism as a political theology of chance: the politics of divination *Pal Commun 1*:1–9

Rawls J (1971) *A theory of justice*. Oxford University Press, Oxford

Rawls J (1993) *Political liberalism*. Columbia University Press, New York

Rawls J (1999) *The Law of Peoples, with "The Idea of Public Reason Revisited"*. Harvard University Press, Cambridge, Massachusetts. London, England

Read J (2009) A genealogy of homo-economicus: neoliberalism and the production of subjectivity *Foucault Stud* 6:35–36

Rhodes M (1996) Globalization and west European welfare states: a critical review of recent debates *J Eur Soc Policy* 6(4):305–27

Rousseau JJ (1762) *Du Contrat Social*. Gallimard, Paris, [1964]

Schneider J, Schneider P (2007) Mafias. In: Nugent D, Vincent J (eds) *A Companion to Anthropology of Politics*. Blackwell Publishers, Malden, Oxford, pp 303–317

Schmidt V, Thatcher M (2014) Why are neoliberal ideas so resilient in Europe's political economy? *Crit Policy Stud 8*:340–347

Sieyes EJ ([1789] (1989)) *Qu'est ce que le Tiers État, Precedé de L'Essai sur les Privilèges*. PUF, Paris

Stiglitz J (2013) *The price of inequality*. Penguin Books, London

Stiglitz J (2016) *The Euro: And Its Threat to the Future of Europe Price of Inequality*. Allen Lane, Penguin Books, London

Supiot A (2013) The public-private relation in the context of today's refeudalization *I CON 11*(1):129–145

Szalai E (2017) Refeudalization. Corvinus *J Sociol Soc Policy 8*(2):3–24

Taylor (1979) What' Wrong With NegativeLiberty? In: Goodin R, Pettit P (eds) *Contemporary Political Philosophy*. Blackwell Publishers, Oxford, pp 418–428

Wacquant L (1999) How penal common sense comes to Europeans *Eur Soc 1* (3):319–352

Walley C (2017) Trump's election and the "white working class": What we missed *Am Ethnol 44*(2):231–236

Weber M (1978) *Economy and society: an outline of interptretative sociology*. University of California Press, Berkeley, Los Angeles

Weyland K (1999) Neoliberal populism in Latin America and Eastern Europe *Comp Polit 31*(4):379–401

Weyland K (2001) Clarifying a contested concept. populism in the study of Latin American Politics *Comp Polit 34*(1):1–22

Whiteside H (2016) Neoliberalism as austerity: the theory, practice, and purpose of fiscal restraint since 1970s. In: Springer S, Birch K, MacLeavy J (eds) *The Handbook of Neoliberalism*. Routledge, London, New York, pp 361–369

Williams H (1994) Kant on the social contract. In: Boucher D, Kelly P (eds) *The Social Contract from Hobbes to Rawls*. Routledge, London, New York, pp 132–146

Wilson K (2012) *Race, racism, and development: interrogating history, discourse and practice. Zed Books*, London, New York

III.
Emotions, Embodiment and Agency

Paradoxes of Emotional Life: Second-Order Emotions

António de Castro Caeiro

de Castro Caeiro, A. (2022). Paradoxes of emotional life: Second-order emotions.
***Philosophies*, *109*(7(5)), 1-23. [109].**
https://doi.org/10.3390/philosophies7050109

1. Some Puzzling Questions

(1) An emotion can be located deep inside the mind. (2) Emotions are inner reactions caused by events situated in the outside world. (3) Emotional reality is indirectly expressed by metaphors. (4) We really need to feel what is up and what is going on to get emotional.[1] One can easily validate these theses unpreparedly. However, on closer inspection, the inside-outside criterion to locate emotions is not very accurate. In the opening line of Wallace Stevens' poem "The house was quiet and the world was calm" , it is the house that is quiet and the world which is calm.[2] Quietness and calmness really exist in the outside world, not inside my head or, more enigmatically, in neurons or synapses. When one experiences a "summer night" one does not need to know anything about neurology or neurophysiology.[3]

Quiet and calm can transfigure an entire evening, but is something inside my mind going outside of my mind to metamorphosize the external world? The thesis presented in (2) deepens (1). Can we establish a causal relation between an object (a thing, a person, a landscape, an event) in the outside world and an emotion inside ourselves?[4] The same object does not always cause the same emotions. Emotions have different shades and several degrees of intensity. Different objects can cause the same emotion in us. At different times, the same object can cause different emotions in the same person or the same emotion in different persons. The same object can even cause a deep emotion in someone and go completely unnoticed by someone else. The thesis in (3) implies that we understand things and persons, allowing us to shuttle between sense and reference, between meaning and facts. There is more about this further down in this paper.

The literal objective reality of facts can be expressed in figurative language. We can export our subjectivity to the outside world. We can

[1] Cf. Jensen and Wallace (2015) on facing emotions.

[2] On experiencing *home*, cf. Purton (2012).

[3] On how different traditions can come together, cf. Elpidorou (2013).

[4] On object emoticons, cf. Martha and Miller (2016).

also import objectivity into our inner world. We already live inside an atmosphere of meaning that admits these apparently disparate and irreconcilable languages. Talking about "quiet house", "calm world", "thinking cigarettes", "cosy rooms", "inhospitable cities" is possible because objective real facts and subjective figurative meanings are different ways of expressing the same reality, i.e., life. Is it possible as (4) states, that there may be emotions that are not felt now while being present. How come? This seems contradictory and paradoxical. How can it be that an emotion is and is not simultaneously present? However, is it not true that there are emotions and sensations that may be building up now, as we have an experience, without being felt? Looking back at any given past moment of our lives, we can feel the emotions we went through without having felt them when they occurred. How is it possible that remembering our high school time we feel all those emotions, feelings, sensations and dispositions that we must have felt then, but did not identify as we carried on with life at that time. Even now, at this very moment, emotions are being formed in a dimension in which we do not feel them. Emotions are not as clear-cut now as they will be sometime next year, perhaps while visiting the same place. Present emotions will arise in the near or distant future as the emotions we are "getting through" now without feeling them.

Life happens in an emotional environment. Emotion is neither interior nor exterior, it is neither objective nor subjective, or it is both subjective and objective. An emotion can be cause and effect. It can be present and not be felt. It can be absent and be felt. It can be associated with an object or a person. It can happen on its own, without any apparent object. Emotions are both private and public, personal, and collective.[5] Even though they can happen momentarily, the impression they make can last forever. Emotions are the ways we feel about life and anything that happens to us. There is, therefore, a spectrum of infinite possibilities of emotions. There are levels of depth. In vol. 20/30, *The Fundamental Concepts of Metaphysics*, Heidegger isolates one fundamental attunement–boredom at three completely different levels of depth. We will follow his introduction to the phenomenon, where

[5] On the effects of boredom collectively, cf. De Lauri (2014). On boredom as an ancient mood: *en têi skholêi*, cf. Bruss (2012).

Heidegger discusses these paradoxical theses. We will try to reach a more comprehensive understanding of the possibility of second-order emotions (4)[6].

[6] We will follow Heidegger's interpretation of the emotional life or attunement (Stimmung). Thence my disclaimer to the Heideggerian: I translate Stimmung as "emotion" rather than "mood" or "attunement" because I want other readers to get to one aspect of the phenomenon at stake. To the non-Heideggerian readers, I say give it a chance. I have been learning a lot about emotional life, studying Heidegger. The phenomenon which Heidegger analyses was identified by Aristotle throughout the corpus aristotelicum as *pathos* but also as *diathesis* (disposition) and *hexis* (condition or way of being (*ekhein* + adv.)). Cf.: Arist. *Metaph*. 1022b1-3: "Diathesis legetai tou ekhontos merê taxis ê kata topon ê kata dynamin ê kat' eidos: thesin gar dei tina einai, hôsper kai tounoma dêloi hê diathesis. (Disposition means arrangement of that which has parts, either in space or in potentiality or in form. It must be a kind of position, as indeed is clear from the word, disposition. Tredennick, 1933.)". In his *Nicomachean Ethics* Aristotle says that the basis of all ethical phenomena are *pathe* (affects, emotions)."Epei oun ta en têi psykhêi ginomena tria esti, pathê, dynameis, hexeis, toutôn an ti eiê hê aretê. Legô de pathê men epithymian orgên phobon, tharsos phthonon kharan philian misos pothos zêlon eleon, holôs hois hepetai hêdonê ê lupê (There are tree kinds of phenomena generated in our mind (*psykhê*): 1. An emotion (*pathos*), 2. A potentiality (*dynamis*), 3. A disposition (*hexis*). Excellence (*aretê*) must, therefore, be one of these three things. By the emotions (*pathê*), I mean desire, anger, fear, confidence, envy, joy, friendship, hatred, longing, jealousy, pity; and generally those states of consciousness which are accompanied by pleasure or pain. Rackham, 1934)." Although the philosophical tradition usually translates the word *pathos* as "affectus" and "emotions", the analyses in the *Nicomachean Ethics* and *Rhetoric* make clear that there is a much more complex dimension to the phenomenon than the linear and superficial one. For Aristotle, it is settled that 'pathê' are forms of perception: "I define *pathê*, on the other hand, as to be such phenomena as wrath, fear, shame, desire, namely such that in general are followed up by a sense of pleasure or pain [unleashed by and] in themselves (legô de pathê men ta toiauta, thumon phobon aidô epithumian, holôs hois hepetai hôs epi to polu hê aisthêtikê hêdonê ê lupê kath' hauta.)" Arist. *EE,* 1220b14-20. The mode of detecting: something sweet (*hedu*) or bitter (*luperon*) implies intentional and emotional changes that promote going after (*dioxis*) an object or running away from (*phugê*) an object. The perceptual opening (*aisthêsis*) to a real object is insufficient to "get" what in that real object brings pleasure or pain. The emotional opening clearly exceeds the reality of a thing. A brown pyramidal object is an uninteresting but clearly objective description of my favourite chocolate, Toblerone. A high-pitched sound is only frightening when it comes from the dentist's drill, for instance. Aristotle seeks to show that for every *pathos* there must be a *dynamis* as its condition of possibility. Without a *dynamis* we would not go through an emotional situation. The affective potential is waiting for the actual encounter with an object to excite an emotional response. In respect of the same content we can see opposite reactions in people. Maybe some people experience some sort of emotion

Heidegger is a name for a phenomenological operator. I put under scrutiny a set of phenomena that Heidegger analyses. I do not intend just to churn out a textual exegesis of Heidegger. The aim is to isolate the specific form of phenomenological openness to the emotional dimension of life. By understanding the situation in which we find ourselves every time already as an emotional situation, it is possible to access phenomena that lie in the same constellation as those that Heidegger presents. They are not, however, the same. If, on the one hand, the idea is to open up the emotional dimension in order to access deep emotional phenomena, on the other, the purpose is to show how the structuring depth of the emotional level makes this same opening possible. The emotional level that we access in an incipient way when we try to isolate first-order boredom may be the same one that is being anonymously and sub-consciously retroactively projected to make those inaugural steps possible. Or not. We may not recognise the deep level that is retroactively projecting itself as resulting in our interest in analysing emotions. For example, our philosophical or psychological curiosity to perceive phenomena such as boredom, fear, anguish, melancholy, and so on is the subconscious being of the emotions. The very being of the emotions is the background that articulates each emotion. The whole emotional level may be a horizon that we know alongside the cognitive and voluntary levels. The being of emotions may be the background that articulates each emotion. The whole emotional level may be a horizon that we know alongside the cognitive level and the volitional level, but

and others do not. If I can sense fear, I am only afraid in a concrete frightening situation. I can be afraid of running away from some situation, though. I can be afraid of feeling shame. In that circumstance there are levels of reaction, response and behaviour that are at play. I cannot avoid feeling fear. But I can aptly respond to it through action. The *hexis* or way of being brave is based upon my resistance to fear. A courageous response allows me to overcome the pain I feel. It all begins with the condition of possibility of being able to have an emotion, sc.: fear, but also to not react to it naturally, i.e., fleeing from fear, but enduring it and waiting for the positive outcome of my action. The comparison between *hexis* and *diathesis* widens the field of understanding of emotion. According to Arist. *Metaph*. 1022b1-3, *diathesis* changes the place, the organization of the place, the appearance of the place in which I find myself. It is not only the noise of the drill that causes me to fear: it is the consulting room in that building, in that neighbourhood of the city of Lisbon, and particularly on the day of the consultation, that appears metamorphosed by the fear I have of the pain felt in the tooth when the nerve is stricken.

we do not understand the unity of meaning of the emotional neither its relation to cognition and volition, nor the possible subconscious or unconscious relations, the short-term reach of an eruptive emotion and the long-term, existential, reach of emotions, the kind of emotions that are definitive and have ontological characteristics that are life defining: the phenomenon of interest, the possibility of love, the passion for truth, the religious scruple, our exposure to the sublime, and so forth and so on. Heidegger is a name for such a phenomenological operator. The examples are not just instantiations of an emotional form, but correspond to experiences in the first person singular. Now, the reader is also a first-person singular and will be able to find the phenomenal basis for understanding what is at stake each time. The balance between the presentation of an autobiographical testimony, an interpretation of an example given by an author who may also be an autobiographical testimony or just an example, is difficult. Moreover, while thinking of the reader, I cannot let him be confronted with auto or heterobiographical examples without him thinking of his own personal experiences. On the level of the emotional example, the private character can only be "destroyed", "deconstructed" and broken down if it has occurred to a person. To identify an experience, we must already have been in a similar situation. Sometimes it may happen that we are in denial of some emotional experience. We may think that we have been in the same situation, or in similar (but not quite the same) situations. Even the protagonist of existence, which is each person in their existence, has very diverse ways of being in the same situation: at a birthday party or on Christmas Eve, with the same people, there are experiences so disparate that they seem to be the experiences of completely different people and yet we are the same "person", it is me that has been there all along. Yet it seems to me that I am almost a different person at each birthday party or on each Christmas Eve.

On the other hand, the formalisation of an abstract thesis – "formally the emotional experience is different from person to person" – does not allow us to understand in a concrete way what this difference from person to person consists of, or what this difference is in various moments of the same person's life. Even if one recognises the same kind of emotional experience, the same form, circumstances make the difference. Even this thesis that the same emotion is different at various levels of depth is experienced differently by several people or by the same person over a lifetime; maybe an abstract thesis cannot not lead us emotionally to the emotional experience of this diversity. Abstract theses about emotions can be understandable if, and only if, the reader has the key to their de- formalisation. Now, the hypothesis that there are different types of otherness requires the passage from the abstract to the concrete, from the impersonal to the personal, from the public to the private. Nothing can facilitate understanding more than "invoking" situations that we went through or opening ourselves up to the possibility of having the experience that is being proposed to us as the only one that demonstrates with the force of evidence what is at stake. No one will know what boredom is unless they have experienced some degree of boredom.

Emotion is a tone, or cadence, that vibrates deep in our lives. It has a real function. No emotion is blind. It offers an understanding of something about something, about someone, about ourselves. Even without our necessarily doing anything about it. We simply feel the power of emotions as feelings articulated in language and interpretation. Every emotion has a script, if we can call it that. We interact with emotions because we do not unequivocally understand what they mean. We just know that they mean something or tend to understand emotions as phenomena that are trying to say something to us. They mean something. The atmosphere we live in, our most intrinsic environment is emotional.

The normal understanding of emotions interprets each emotion (1) as the mental effect of an 'exciting' agent in the outside world; (2) as a mental phenomenon within the *psykhê*;

(3) as a metaphorical expression. Furthermore, (4) to become "emotional", emotions must rise to the surface of consciousness. We suggest that this is the other way around. In fact, we will first begin with the reversal of (4). We suspect that emotions may be present, conditioning our lives at any given moment, without "declaring" their presence to us. On the other hand, some emotions do make their presence known but are superficial, they vanish as soon as they show up. Emotions constitute the ultimate frontier of our conscious life. They admit depth. If so, can

we go deep enough to get a glimpse of this emotional dimension where meaning, sense, and understanding of what is going on may be available? If so, do theses 1–3 collapse? Heidegger's analysis of boredom in its deep fundamental dimension will allow us to understand the different layers and degrees of emotional experience. We know too well, as a matter of fact, that we do experience boredom in some situations. This empirical level is enough for a first characterisation of this emotion on a superficial level. From then on, we will try to understand how we can get to a deepening of boredom, specifically, and of other emotions or dispositions on dimensions that at first glance, and for the most part, are removed from our normal everyday experience.

So, what is decisive for us is to understand how boredom can emerge as a fundamental deep emotion. Can boredom, like any other fundamental emotion, be happening without our having a perception of it. Let us therefore try to understand (1) how the cause-effect relationship fails; (2) how boredom is so external that it characterises a being in the external world (a person, a situation, a moment in our lives); and (3) whether the metaphor is already the primordial expression of our relationship to life (towards the world, others, ourselves)? Does this mean that even facts are already expressed through metaphors and analogies? It may happen that meaning antecedes facts, and sense overshadows reference. Therefore, metaphors express the a priori emotional grounding of all our life experiences. Our take is, therefore, that there are levels of emotional experience, ranging from concrete situations full of emotion to dimensions we can only reach after sufficient time has elapsed for us to understand emotions that were present then at that past time.[7] Does this mean that we are going through emotional layers in our present moment of life that we can only experience as unambiguously emotional in the future? How can we have the foresight of the emotional forecasting of our future?

[7] On constructing emotional past life cf. Jensen and Wallace (2015).

2. Boredom as a Fundamental Emotion

How, and why, does Heidegger suggest that boredom is the emotion that attunes our life? Out of all our emotional experiences assessed as positive and negative, why does he stress deep boredom? Why not, say, love? We do identify things, people, situations, phases, and periods in our life as boring, but how is it that boredom lies as a deep fundamental emotional disposition at the bottom of all beings? So, we understand that some beings are boring. Sometimes we are boring, but can we say that we react to an anonymous presence of boredom all the time in our lives? Can a fundamental emotional disposition be constitutive without showing up all the time? Can we be bored without "feeling" boredom? Would it be possible to fill up our lives with all sorts of activities we enjoy performing all day long and be bored? This amounts to feeling emotions (enjoyment, pleasure, fullness) completely different from boredom, but let us not rush in dismissing Heidegger's thesis as not making sense yet. We do have the perception of moments in time that are interpreted as setbacks and delays. We get stuck in moments. Time seems to freeze. Sometimes, we do experience feelings of emptiness, situations without any meaning. Are these moments in which boredom manifests itself enough for us to go deeper into the depths of that dimension where emotions lie without declaring their presence to us? Can we go down to such depths? How? Could boredom emotionally attune our personal and collective lives? Is Heidegger's analysis circumscribed to a time, almost one hundred years ago, and should it be dismissed on historical sociological or even psychological grounds? Do not our lives in the 21st century seem so much better and full? Or can boredom still be alive getting us both at a personal and collective level? Can we vibrate with joy, loving what we do, and at the same time be bored to death? The diagnosis of our current situation stems from a sensation and, as such, it is a short-lived phenomenon.

> "Das Ganze ist eine Sensation, und das heißt immer eine uneingestandene und doch wieder scheinbare Beruhigung, wenn auch nur literarischer Art und von charakteristischer Kurzlebigkeit. (Everything-in-its-entirety [das Ganze]

> is a sensation [Sensation], and that always means an unacknowledged and yet ap- parent serenity, even if it is only of a literary nature and characteristically short-lived.)".[8]

We get into a situation in which impressions caused by something or someone are felt, leaving us in a certain state of mind. Can something or someone be the cause of the sensation we feel? What if the sensation is caused by us and then spreads to everything, over a period? For example, on a Sunday afternoon, even for a fleeting moment, everything seems to be boring. This sensation disappears as quickly as it arises. We do not know where it came from or where it went. We do not know why boredom showed up, but we know that an entire city can be engulfed by this feeling of ennui that pervades everything. Is boredom asleep just to wake up on a Sunday afternoon?

Emotions are at the basis of the philosophy of culture. Its aim is a diagnosis of culture (Kulturdiagnostik).[9] Now, every diagnosis presupposes a prognosis, "it can constitute and become a prognosis (zur Prognose aus- und umbildet)".[10] The philosophy of culture seems to have good intentions. Knowledge of our personal and collective past, nationally and worldwide, seems to allow us to get to know our present better so that we can understand where we are going, but words such as "diagnosis" and "prognosis" seem to suggest a civilisational disease. Are we in danger as a species? When one talks about survival, the planet's resources are not the case in point. The survival at stake is none other than that of existence itself based on the emotional dimension where all our aspirations and expectations take place.

Heidegger's perspective is different than that of the cognitivists. His philosophy is much closer to Kierkegaard and Nietzsche than to that of the philosophers of emotions. For Heidegger's commentators, he belongs to that 19th-century tradition which operated the inversion of the hierarchy attributed to the psychological acts. Therefore, while tradition

[8] Heidegger: *GA 20/30* (112).

[9] *Ibid.*

[10] *Ibid.*

has privileged representational acts (logos) over volitional (ethos) and emotional (*pathê*) acts – a thesis popularised by the Stoics and upon which even Kant bases his philosophy, Heidegger reversed that order. However, he does not say that first is the emotional, then the volitional, and only thirdly the logical or representational. He says what Scheler and Husserl had already intended. He bestows a noetic dignity upon synthetic acts of the mind (synthetischen Gemütsakte). With phenomenology, love ceases to be blind. Heidegger seeks to situate the revelation of being, the truth of being, on a plane prior to that of representation itself. The question of the meaning of being is posed by the discovery of dispositions grounded on a plane of openness and closure that is that of the situation in which each of us finds ourselves. Reflection and self-perception can completely block my access to myself. On the other hand, I can absolutely wish to hide from myself and not succeed, because my most irrational fears find me, the most sordid dreams find me, and the most abject thoughts find me.

Therefore, the emotional is not mental content, nor a *noêma* or a "representational" content, but the reality as such. I do not love the mental content of my mother. I love my mother. The relation between myself and my mother is already emotional. Why duplicate the objects? What Heidegger seems to do is to get rid of those presuppositions that do not help but rather impair any comprehension of emotions of what is already happening.

The majority of Heidegger's analyses from the point of view of the hermeneutic tradition, even when they stress the affective or dispositional turn, consider emotion as a regional aspect of an ontology. It is not clear, however, how one releases a personal emotion – an affective crisis with emotional impact, for example – in order to interpret it in a philosophical dimension. Even when one goes through a heart-rending experience of romantic (*erôs*) or religious (*agapê*) love, there is an enormous difficulty in understanding how others can go through the same thing or how these feelings could both have been at the basis of all the stories of romantic love and lead the Christ to the cross. From the point of view of the non-continental tradition, Heidegger is interpreted, sometimes sympathetically, as advocating fundamental aspects of pragmatic philosophy and clearly adding to the emotional plane, which can be blind and mechanical, or merely a "cognitive" (instinctive) aspect. Now, Heidegger criticises the primacy of the cognitive as access to the world. In his view, in-Sein is affective and not cognitive. This aspect of openness and access to the totality of being is obscured by both the hermeneutic and the analytic or even pragmatic traditions in their approach to the emotions. Heidegger's philosophy is a "Stimmung". It is from the interpretation of *thauma* and *thaumazein*, or philosophy as *erôs* and as *pathos*, simple and absolute, that Heidegger must be approached. In this same way, the erotic experience in the platonic sense of the term is an experience of the maximisation of our personal being (in our relationship with things, with others and with ourselves), because we have access to an absolute, maximum, superlative version of ourselves. It is the superlative, emotionally open, version of ourselves that is looking at us now, here in the present, from the future. The absolute exponentiation of myself puts enormous pressure on the version of each of us now. This tension results from an ulterior version, possible but effective. Each of us projects ourselves towards this superlative possibility of our own self, open ourselves to the other, seek another to love, seek God in religion, seek the sublime in art. Thus, for Heidegger, being and the truth of its meaning, the problems of philosophy, open up in the affective tone of an emotion. Being opens up in moments of truth and revelation emotionally. The emotional plane is the agent of *alêtheia*. The experience of being is always emotional and affective. It happens to us on the plane of everyday life because such visitation is possible. What we need are eyes to see it. There is thus always an exposure and a vulnerability in everyday life to this ecstatic, existential dimension of emotion. Existence in Heidegger means: to be [continuously] going outward [from within], to manifest oneself. However, if the temporal organisation can naively be: past, present, future, in Heidegger it is inverse: future as possibility (possible or simply impossible) (Entwurf), present (Verfallenheit), past (Gewesenheit). It is the future that goes out of itself and approaches us in the present, in a "movement" of ever smaller and smaller inflows and ever larger and larger outflows. The flow of existential time and the structurally temporal disposition is an ever-smaller influx of possibilities coming from the future, and an ever-larger flow of lost possibilities in the past. Centring Heidegger's analysis on these aspects leaves Gadamer's hermeneutic interpretations in *Wahrheit und Methode* – the matrix text of Hermeneutics – far short of what Heidegger intends. Damasio's analysis of the emotions, for example, is blind to the total and ontological dimension of the problem of being and the truth of being that Heidegger seeks to pose.[11]

This is the reason why Heidegger speaks of the awakening of a (not "the") fundamental disposition. What if boredom is what emotionally, and therefore existentially, constitutes our lives? "It is precisely this the reason why we are striving to awaken a fundamental emotion (gerade wenn und weil wir die Weckung einer Grundstimmung erstreben)".[12] However, awakening is different from observing. "There is a theoretical difference between observing for cognitive purposes our spiritual situation and awakening a fundamental emotion (Es besteht ein theoretischer Unterschied zwischen der Darstellung der geistigen

[11] For the contemporary debate, see Jensen and Wallace (2015) and Gilje (2016). Kriterium. Specifically for debating the works of Robert Solomon and Matthew Ratcliffe, cf. Elpidorou (2013). The paralell drawn from the philological point of view shows clearly how different the output of different traditions can be almost juxtaposed. But the methodological approach in phenomenology is almost never taken into account (Malabou 2019, Solomon 1988). Most interesting for our purposes is Capobianco (1993). Capobianco presents Jung's take on the "unconscious" as "an intelligent, transpersonal structure", which "allows opposites to "happen" together and, thus, is irreducible to consciousness". Consciousness (ego) and unconsciousness "are not reducible one to the other" but "are nevertheless mutually dependent". "The unconscious maintains a primacy over consciousness, … can "fascinate" and "overpower" the ego; even as the unconscious maintains primacy over "subconscious" and "consciousness" Capobianco (1993). Capobianco argues that Heidegger's take on ego must be understandble from his redefinition of subjectivity as Dasein, meaning that there are multiple ways in which the "I" "is". But it never gets to the bottom line of Heidegger's subjectivity as anonymous and deep. In GA vol. 20, Heidegger says that Descartes "discovered" the "ego" but has forgotten to say anything about the "sum". Implicitly in the "sum" there is a "moribundus": I am means, I'm about to die as long as I live. "The appropriate statement pertaining to Dasein in its being would have to be sum moribundus ["I am in dying"], moribundus not as someone gravely ill or wounded, but insofar as I am, I am moribundus. The MORIBUNDUS first gives the SUM its sense [Sinn]. (GA 20: 437f/317) [Translated by Theodore Kisiel]. We shouldn't forget that "the task of the philosophers (der Philosophen Geschäft)" presented in *Being and Time* is the same in Kant's anthropology: i.e.: explicitly to "discover" "the hidden judgments of common reason (die geheimen Urteile der gemeinen Vernunft)." Heidegger (1927).

[12] Heidegger: *GA29/30* (113).

Lage und der Weckung einer Grundstimmung.)".[13] This means that our situation may be symptomless. The emotional symptoms can be ambiguous. Do we need for the main character of one's life, oneself, to get "sick" with boredom? For if we do not feel boredom, we need to stimulate its presence. We should at least try to remember any boring past situation and dive into it to understand what has happened.

What is valid for boredom is valid for all emotions. All emotions have a layout linking their unmistakable surface eruption and their deep existence. In that sense, "we would not be allowed to ask where are we? We should rather ask how it is with us? (dürfen wir nicht fragen: wo stehen wir?, sondern müssen fragen: wie steht es mit uns?)".[14] How can we delve into the moments in time when we were bored, when time had stopped and everything was empty? At that time, we cannot ask: "I wonder if what happens to us after all is that there is a deep boredom in the depths of our existence that, like a silent fog, pushes and pulls us wherever it goes? (Ist es am Ende so mit uns, daß eine tiefe Langeweile in den Abgründen des Daseins wie ein schweigender Nebel hin- und herzieht?)".[15]

What do people, things, situations have in common when they all become boring? What do books, shows, evenings, holidays have in common to be or become agents of boredom? Is it because "we ourselves get bored because we become bored with ourselves? But must the human being himself become bored with himself? Why is this so? (Etwa weil wir selbst uns, uns selbst, langweilig geworden sind? Der Mensch selbst sollte sich selbst langweilig geworden sein? Warum das?)".[16]

We need to cast off all theories. It is through direct observation that we know what happens when we are bored: time stops, and one feels emptiness. Does the depth of a disposition directly correspond to the power with which it appears and the scope it has? We tend to think that the stronger the emotions are the more real they are, but could an emotion, despite not being felt, exist and exert pressure upon our life,

[13] Heidegger: *GA29/30* (114).

[14] *Ibid.*

[15] Heidegger: *GA29/30* (115).

[16] *Ibid.*

even without our being aware of it? How do we know whether this emotional dimension really and effectively exists if it is not manifest? Perhaps we can understand its anonymous presence through our way of being, acting, in our attitudes and behaviour. We avoid certain situations because we think they will bore us, but how is future boredom captured? It is undeniable that there are people and situations that cause us boredom. What is manifested in this direct contact with boredom is that time slows down and everything seems empty and superficial. We can now find stimuli that move us or interest us.

> "We had just asked if it happens to us after all that a profound boredom moves back and forth in the abysses of existence like a silent fog. (Wir haben nur gefragt: Ist es am Ende so mit uns, daß eine tiefe Langeweile in den Abgründen des Daseins wie ein schweigender Nebel hin- und herzieht?)".[17]
> "What does it mean: is boredom problematic for us? First, formally, it says as much as this: we do not know if it conditions us emotionally now. (Was heißt: die Langeweile ist für uns fraglich? Zunächst sagt das formal soviel: Wir wissen nicht, ob sie uns durchstimmt oder nicht.)".[18]

(1) This emotional feeling of boredom has no reality. Yet, it may be having an effect. How can a phenomenon like a disease exist causing effects without symptoms? Its manifes- tation seems to be fully armoured to us. (2) Boredom as emotion is metaphorically described as a haze. Unlike a thick fog, it lifts and disperses. It is concentrated in certain places and very light and tenuous in others. It sways as the wind blows. "In the end, we do not want to know anything about that emotion, but we are constantly trying to avoid it. (Wir wollen am Ende nicht von ihr wissen, sondern suchen ihr ständig zu entgehen.)".[19]

Like all fundamental emotions, profound boredom can exist in a deep, unfathomable dimension. However, it can erupt from time to

[17] Heidegger: *GA29/30* (117).

[18] *Ibid*.

[19] *Ibid*.

time in episodic moments. We may even not realise that it is boredom. We do not even have a name for this outbreak. Thus, it may be that we know what boredom is because it has already been present in our lives with its devastating power. Based on those past experiences, we do not want to know anything about it, let alone awaken this emotion in order to interpret its deep wisdom. Boredom is not a pleasant feeling. Maybe all deep dispositions have this uncanniness about them, we do not want them to break free from wherever they are kept. However, we understand that we are continually reacting to them, anticipating their presence. We may try to always be busy, have things to do, close ourselves off from the manifestation of those emotions that are coming from that deep dimension of existence.

> "How to escape boredom in which, as we say, time becomes long? We are simply striving so much, consciously or unconsciously, to pass the time, that we welcome the most important and essential occupations, just because they fill our time. Who wants to deny that? But then, is it still necessary to ascertain that this boredom is there? (Wie entgehen wir der Langeweile, in der uns, wie wir selbst sagen, die Zeit lang wird? Einfach so, daß wir jederzeit, ob bewußt oder unbewußt, bemüht sind, uns die Zeit zu vertreiben, daß wir wichtigste und wesentlichste Beschäftigungen begrüßen, schon allein, damit sie uns die Zeit ausfüllen. Wer will das leugnen? Bedarf es dann aber erst noch der Feststellung, daß diese Langeweile da ist?)".[20]

We need to kill time, we want "to kill time", we want time to pass. We tend to be running away from boring situations, trying to escape them when they are there, but we know that boredom can come at any time. How come? What grounds this knowledge about this emotion? We know very well that boredom can always come again. Like all deep and profound emotions, they are asleep for the most part of our lives. We all know what depression, melancholy and sadness are when they

[20] Heidegger: *GA29/30* (118).

are present, but we also know that, somehow, they can and will show up again in our lives. They vanished but they did not go away forever. We know they can come back. We know this because in the past they had disappeared but then reappeared again. Every episode of deep emotional experience brings with it interpretive intelligibility. Emotions are intelligible. They allow us to understand what is going on with us when they manifest. Even if they do not release a full-fledged knowledge of ourselves, we know all deep emotions have something to say about us. It is from the future that deep emotions come to us. When in a deep phase of depression, we may get out of bed without feeling the presence of depression. We have our breakfast; take a shower. We might wonder why depression has not yet arrived. It feels like we are anaesthetised. However, when dealing with anxiety and depression, we know it is only a matter of time, and then it comes. Anxiety or boredom attacks us. What happens between the moment we get out of bed and the moment we feel the presence of anxiety? Is this absence of feeling a depressive emotion the same as when we do not feel boredom? Is it different?[21]

"But what does it mean: we expel and drive boredom away? We always make boredom fall asleep. (Was heißt das aber: wir vertreiben und verscheuchen die Langeweile? Wir bringen sie ständig zum Einschlafen.)".[22] Killing time, occupying time, filling up time has a clear and unambiguous meaning: to anaesthetise us against the uncomfortable presence of boredom. We may even come across tasks, occupations, that have become mechanical or automatic as a pragmatic reaction to the presence of boredom. When we feel boredom kicking in, we try to get busy, we keep ourselves occupied, as a sort of self- conceived occupational therapy, but does free time expose itself to boredom? Is boredom already ever-present, waiting for the right moment to show up? With what intention?

"We 'know' – what a remarkable thing to know it – that boredom can always return at any moment. (Wir "wissen" – in einem merkwürdigen Wissen –, daß sie doch jederzeit wiederkommen kann)".[23] This knowledge

[21] Boredom and anxiety have the same common bottomless ground. Cf. Sheets-Johnstone (2015).

[22] *Ibid.*

[23] *Ibid.*

of how it is with us (in relation to an emotion) opens up the possibility of being affected by profound dispositions. The way emotions appear and disappear, rise and fall, hover above us and harass us is intrinsic to this phenomenon. We avoid boredom and situations that we think will bore us. However, where does boredom as such come from? The same question can be asked regarding anxiety, anguish, fear, melancholy, and sadness. No emotion of this deep kind disappears forever.

> "So, boredom is already there then. We try to kick it out. We try to put it to sleep. We do not want to know anything about it. This does not mean that we do not want to be aware of it, it means, rather: we do not want to keep it awake – this boredom that is after all already awake and with its eyes wide open – even if absolutely keeping its distance – looking into our Da-sein from the outside and penetrating and tunin g us with its look. (Also, ist sie schon da. Wir verscheuchen sie. Wir bringen sie zum Einschlafen. Wir wollen von ihr nichts wissen. Das heißt ja gar nicht: wir wollen kein Bewußtsein von ihr haben, sondern es heißt: wir wollen sie nicht wach sein lassen – sie, die am Ende doch schon wach ist und mit offenem Auge – wenn auch ganz aus der Ferne – in unser Da-sein hereinblickt und mit diesem Blick uns schon durchdringt und durchstimmt.)".[24]

3. Emotional Stalking

Heidegger avoids talking about consciousness or unconsciousness as far as our attune- ment to emotions is concerned. Being aware of an emotion is totally different from waking it up or not letting emotions slip out of our minds, or even putting them to sleep. We say we do not want to think about emotions whose content causes us anxiety or suffering. Yet, that is all we think about. It imposes itself. The opposite can also happen. Sometimes we would like to "feel again" what we had felt in a certain

[24] *Ibid.*

situation. However, those past feelings and emotions do not show up again to be "lived" by us. To trigger a past emotion is to tune into it, but we can wake up emotions that were not even felt in past situations. There is, however, a faint vibration of this past manifestation in the present. We cling to this emotional thread so that past emotions and feelings can fully manifest as they were. By so doing we try not to let sentimental references to the past "die" entirely or at least not let them fall asleep. There are emotions, affects, feelings that exist as living characters. Some are like people dear to us whom we love. There are also sinister characters who inhabit the attic or basement of our lives. They are clandestine but we know perfectly well that they cohabit with us. Usually, we do not come across them. Yet, we feel that these emotional characters follow us at every step. They are constantly stalking our existences. They know everything about us, what we do, what we think, who we are, how we are. Feeling the gaze of the emotional disposition from the outside, at a distance, affects us. This is how the look of emotional depth feels like. We feel vulnerable to that gaze of emotions. We are transfigured by the look of emotional damage. Now, it is precisely this transfiguration that we do not want to let fall asleep. We do not want to escape from it. On the contrary, we want to be running towards that emotional feeling that lets us know something about ourselves. All we need to do is strive to keep that emotional depth awake.[25] 2Maybe we can then learn about ourselves from it. Emotions do say "something" about us. They are agents of truth.

> "But if emotions already are awake, then they do not need to be woken. Not really. Awakening this fundamental disposition does not mean waking it up first, but letting it stay awake and preventing it from falling asleep. We can easily infer from this that the task has not become any easier. Perhaps this task has become much more difficult; perhaps because we know that it is always easier to wake someone up with a shock than to prevent them from falling asleep. (Aber wenn sie schon wach ist, dann braucht sie doch auch nicht geweckt zu werden. In der Tat nicht. Das Wecken

[25] On depth, cf. Mendonça (2019).

> dieser Grundstimmung heißt nicht, sie erst wachmachen, sondern wachsein lassen, vor dem Einschlafen behüten. Wir entnehmen hieraus leicht: Die Aufgabe ist nicht leichter geworden. Vielleicht ist diese Aufgabe wesentlich schwieriger, ähnlich wie wir jederzeit erfahren, daß es leichter ist, jemanden durch einen Schock aufzuwecken, als ihn vor dem Einschlafen zu behüten. Doch ob sie schwer oder leicht ist, das ist hier unwesentlich.)".[26]

So, the task is finding the fundamental emotion that, by vibrating, makes us under- stand what goes on with us at the bottom of our lives. This emotional depth is at first shielded. It is seen as always existing vibrantly. It "exists" from afar but close enough to watch our life at every moment. Fundamental emotions and feelings have already surfaced. They came and went, after being alive for a lapse of time. The difficulty, then, is to recapture in some way what feelings let us feel. What happened when some emotions were present, affecting all our life with their presence: the world, others, ourselves. How can we reconstruct the story of an emotion and our dealing with it? How can we resuscitate what it made us feel, the impression it made? How are we to measure its emotional impact? What is the state of mind it left us with? If there still lives in us an emotional hint of what went on, we do not let it fall asleep, even if all that is left of that emotion is only a shred of life. However, it is very hard to know how to keep awake an emotion that wants to go to sleep. Or is it me that somehow wants some emotions to go to sleep?

"The task, then, is not to let boredom go to sleep (Die Langeweile nicht einschlafen zu lassen)".[27] From the outside, this methodological task seems to counteract our tendency in life to thwart and resist motionless moments of time, setbacks, and delays. We always want life to keep on going. Yet we know all too well about those moments when we feel bored and are stuck in them.

When a circumstance of boreom forms, we try to "kill time", to chase away and dissipate the oppressive presence of boredom. We try

[26] *Ibid*.

[27] Heidegger. *GA29/30* (119).

not to allow it to be there. When it comes to pass, we do not want it to stay awake, we try to put boredom to sleep ("die Zeit vertreibt und die Langeweile gerade nicht aufkommen läßt, das heißt, wenn sie kommt, sie verscheucht, sie zum Einschlafen bringt?)".[28] Now, however, "we have to keep boredom awake (Wir sollen sie wachsein lassen.)".[29] We still do not know how, because this move is against the natural tendency of life, which is to try to escape boredom, by finding a way to occupy time, by finding occupations so as not to let life "stand still". So, we chase boredom away. Now we want to do the exact opposite, but how to proceed? Whenever we feel bored, should we just let it endure and try not to move so that boredom grows? When boredom arises, will we have enough presence of mind not to react, to try not to think about something else? However, to think about what is happening to us, does it not disturb the very presence of boredom? Can we, by remembering situations of boredom, bring back the state of mind in which we lived them and thus release the disposition felt in the past so that it can attune us to it?

> "Boredom has a varied multiplicity of figures (Gestalten) that we know all too well in their most diverse disguises and masks (Verschleierungen). When it emerges (auftaucht), it affects us in the blinck of an eye, for a moment, or else tortures and afflicts us for long periods of time. As soon as it appears, it is there, we try to repress it, we strive to expel it from our lives ("Die Langeweile – wer kennt sie nicht, wie sie in den verschiedensten Gestalten und Verschleierungen auftaucht, uns oft nur für Augenblicke befällt, oft auch längere Zeit quält und bedrückt. Wer weiß nicht, daß wir, sobald sie kommt, uns auch schon daran gemacht haben, sie wegzudrücken, und bemüht sind, sie zu vertreiben)".[30]

Are there other forms of boredom so profound that they are not identified as boredom, at least most of the time and at first sight? Is it possible to

[28] *Ibid.*

[29] *Ibid.*

[30] *Ibid.*

go through vibrant situations as opposed to the monotonous cadence of boredom, and still identify them as boring, for example, when having fun or working enthusiastically? Having fun or working hard can be ways of spending time, but they can denounce that we are bored, or otherwise we would not try to escape. Can fun and work be just the surface of boredom?

> "Or is this boredom that we know and of which we now speak in an indeterminate way a mere shadow of true, genuine, authentic boredom? In fact, we ask and continue to ask again and again: is it really happening to us that deep down there is a deep boredom in the abysmal depths of our existence pulling us this way and that to determine our own existence? ("Oder ist diese Langeweile, die wir da so kennen und von der wir jetzt so unbestimmt sprechen, nur ein Schatten der wirklichen? Wir fragten ja und fragen immer wieder: Ist es am Ende so weit mit uns, daß eine tiefe Langeweile in den Abgründen des Daseins wie ein schweigender Nebel hin- und herzieht?)".[31]

There is a relationship between boredom and time. The duration of boredom episodes can be short or long. However, the time of emotions does not make them exceptional. Everything has a beginning, middle and end in time. What makes the relationship between boredom and time so special? Boredom lengthens time. When it happens, there is a metamorphosis of time in us. The time of life, which passes continuously without us always being aware of it, emerges to the surface. Even any superficial manifestation of boredom allows us to understand the essential relationship we have with the time of existence. This relationship with deep time cannot be undone.

> "Boredom (Langeweile) indicates almost palpably, and especially in our German word, a relationship with time, a way in which we situate ourselves in relation to time, a feeling of time (Langeweile... zeigt fast handgreiflich, und

[31] *Ibid.*

> besonders in unserem deutschen Wort, ein Verhältnis zur Zeit, eine Art, wie wir zur Zeit stehen, ein Zeitgefühl. Also führt uns die Langeweile und die Frage nach ihr zum Zeitproblem.)".[32] "...or is it the other way around and boredom only leads us to time, to an understanding of how time vibrates in the depths of existence and of how we can, therefore, 'act' and 'maneuver' alone in our usual superficiality? (Oder ist es umgekehrt, führt uns die Langeweile erst zur Zeit, zum Verstehen essen, wie die Zeit im Grunde des Da-seins schwingt und wir deshalb in unserer gewohnten Oberflächlichkeit allein "handeln" und "lavieren" können?)".[33]

4. The Four Paradoxes

It is now time to return again to where we left off. Starting now from the interpretation of boredom, can we understand how the three schemes for explaining emotions do not apply? (1) The cause-and-effect relationship (das Ursache-Wirkung-Verhältnis); (2) the subjective interiority of emotions (das Innerseelische); (3) the metaphor (Übertragung) as an expression of emotions; and (4) the reductive manifestation of emotion blocking the experience the deep emotional level. Schemes 1–3 do not even apply to the superficial level where emotions show up, but let us apply the three schemes to boredom. The starting point should be to identify the active element of boredom in a boring situation (im Ausgang vom Langweiligen). When we feel we get bored by something, we can identify the active element in something or someone that bored us in each situation. On the other hand, we are bored because we are liable to getting bored. At last, the synchronicity of the active and passive elements produces the whole situation of boredom. As we have seen, once in a situation of boredom time slows down and it comes to a halt. One feels it is never time. Time stands still. We are kept on hold, waiting

[32] Heidegger, *GA29/30*: 120.

[33] *Ibid*.

(das Hinhaltende). The second elemental characteristic is the feeling of emptiness. We are deprived of sense and meaning mainly because in that situation there is nothing telling us anything in order to fulfil our being there (das Leerlassende). In this sense:

> "das Verstehen der Stimmung verlangt von uns am Ende einen Wandel der Grun- dauffassung des Menschen. Die rechtverstandene Stimmung gibt uns erst die Möglichkeit, das Da-sein des Menschen als solches zu fassen. (Understanding of emotions requires of us, after all, a transformation of the fundamental un- derstanding of being human. Emotions correctly understood finally give us the possibility to grasp the Da-sein as such of the human)".[34]

Our relationship to the active element of boredom in an object is a relationship with that boring object and not a relationship to our representation of the emotion "boredom".[35] Heidegger stresses again and again that when we feel bored, it is due to the relationship with an object really existing in the outside world and not with the emotion felt in the stream of consciousness, inside our mind.

> "Langweiliges kennen wir so, weil es in und durch seine Langweiligkeit in uns Langeweile verursacht. (We know the "boring", the active element in a thing, because the quality boredom as such is what in its essence and through itself

[34] Heidegger, GA29/30: 123.

[35] Heidegger stresses different aspects in this situation. He is trying to underline that emotional phenomena are experienced in the "factical life" (Faktizität). Therefore we need to avoid the unprepared interpretation that we are dealing with phenomena inside our minds (*psykhê*). He starts from the quality of "boringness" (Langweiligkeit). Cf.: GA29/30: 123: "So gehen wir schon zu Beginn absichtlich nicht von der Langeweile aus, schon deshalb nicht, weil es dann allzusehr danach aussieht, als wollten wir ein seelisches Erlebnis in unserem Bewußtsein der Analyze unterwerfen." Our underlining. We start with the boringness (Langweiligkeit): "Formal gesprochen ist die Langweiligkeit das, was etwas Langweiliges zu dem macht, was es ist, wenn es langweilend ist. (Formally formulated, the ontological quality of boredom is what makes something boring when it is boring.)." It is the experience of the active element of boredom, that which is boring, that makes any content boring.

> makes us bored)".[36] What can "boredom" be? "Etwas Langweiliges – ein Ding, ein Buch, ein Schauspiel, ein Festakt, aber auch ein Mensch, eine Gesellschaft, aber auch eine Umgebung oder eine Gegend – solch Langweiliges, das ist nicht die Langeweile selbst. (Everything can be boring – a thing, a book, a show, a party, but also a person, a company, but also the surroundings or a region)".[37]

There are three figures of bordom that Heidegger analyses. The concrete experience of boredom can be totally passive. The formulation is in the passive voice and accentuates the boring agent, that which leaves in us the sensation of boredom: "Von Langweiligem werden wir gelangweilt, so daß wir uns dabei langweilen." (We are so bored by the boring that we are bored with it)".[38] There is something or someone, a situation, circumstance, that "causes" an impression on us and leaves us in a boring state. This is being bored "by" (das Gelangweiltwerden durch ein solches Langweiliges). The second form of boredom is reflexive: getting bored by a situation (das Sichlangweilen bei einem solchen). The third is: boredom as such (die Langeweile selbst)".[39] This structural analysis identifies several possible figures of boredom, which go far beyond the subject–object relationship, given their complexity. We can understand the different manifestations of boredom using the passive voice (to be bored and be bored by), the reflexive voice (to be bored with) and the active voice: (to bore, boredom). Boredom is a *nomen agentis* expressing the intransitive verb to bore. By identifying this figure, Heidegger aims at presenting several complex forms of experiencing boredom for which the subject–object relationship and the philosophical classification ideal and real, empirical and rational, are short-sighted categories. In all three forms, the starting point is always the concrete experience of boredom, though. Each figure stresses one aspect of boredom. The same object does not always have to be boring, nor do we always get bored with that same object.

[36] Heidegger, *GA29/30*: 124.

[37] *Ibid*.

[38] *Ibid*.

[39] *Ibid*.

5. Emotions Outside and Things Inside?

> "Die Langeweile ist nicht einfach ein seelisches Erlebnis im Inneren, sondern etwas von ihr, das Langweilende, was das Sichlangweilen entspringen läßt, kommt uns gerade aus den Dingen selbst entgegen. Die Langeweile ist viel eher draußen, sitzt im Langweiligen, und von draußen schleicht sie sich in uns ein. (Boredom is not simply a psychic mental experience inside our minds, but there is something in boredom, the boring itself, which triggers boredom, and which comes to us precisely from the things themselves, in the outside world. The boring element in things comes from the outside and insinuates itself into us".[40]

The boring thing is the thing (Ding) out there: the people themselves, the events, shows, landscapes, regions, dust from the house, temperature, everything that is out there. There is not an object and a boring element in it. It is that thing that is boring. This discovery shifts the centre of gravity in the analysis from the mind to the things themselves. Indeed, boredom is neither in the chemical elementary structure of a spatial-temporally determined and specified thing in its corporeal matter, nor is it found in neurons or synapses as such. That is why the experience of something or someone as boring is the original starting point from which we begin our analytical thinking. It takes place before any theory that one may have about phenomena of this nature, for we have always known what it is like to deal with boring things and persons in boring situations. What is decisive here is the concrete experience of emotions at play. So, we need to know how to deactivate all explanatory theories about emotions and feelings based in science. All discoveries come later than the phenomena. We do not need to wait for science to understand the powerful effect of an emotion. What do we find boring in a person, in a show or in a book? They drag us on and dry us up. The boring exists in the book in its relationship with us, in seeing it as an object and in reading it. The

[40] *Ibid.*

boring drags on and is arid. By dragging on, it slows us down. When we find something as boring, we say that it drags on. Life is put on hold. At the same time, it is arid and dry. It contaminates us. Makes us empty.

> "Langweilig – wir meinen damit: schleppend, öd; es regt nicht an und regt nicht auf, es gibt nichts her, hat uns nichts zu sagen, geht uns nichts an. (Boring: what do we mean by this: it drags on, is arid; neither stimulates nor provokes anything, gives nothing of itself, has nothing to say to us.)".[41]

6. Causal Relation

After the identification of things outside of us, it is important to understand the causal relation between things and us in order for them to be understood as boring. What do they provoke in us? How do they cause us to experience boredom?[42] The boring "ist das, was uns langweilt, also Langeweile verursacht. (The boring is that which bores us, therefore that which occasions boredom.)".[43] "Was das Langweilige in seiner Langweiligkeit ist, können wir doch nur aus der Langeweile verstehen, und nicht umgekehrt. (Only from boredom are we able to understand what the boring is in its boring quality and not the other way around)".[44] Boredom is the meaning from which we are able to identify boring objects there in the real world of life outside ourselves and, conversely, emotions of boredom. The verb "occasion" (Verursachen) can be read as causing an effect, but it never recognises a mechanical or physical causal relationship between an object in the outside world and an emotion inside our minds. It is not like when the temperature drops and the thermometer shows the mercury falling. Or when a billiard ball

[41] Heidegger, *GA29/30:* 126.

[42] Heidegger, *GA29/30:* 124: "Was das alltägliche Sprechen und Verhalten und Urteilen zum Ausdruck bringt".

[43] Heidegger, *GA/2930:* 125.

[44] *Ibid.*

sets another ball in motion, diverts the course of another ball, or stops it.[45] To occasion means to propitiate, to bring together the conditions for something to happen, even though it may not happen.

> "Denn was heißt es, gewisse Dinge und Menschen verursachen in uns Langeweile? Warum gerade diese Dinge und jener Mensch, diese Gegend und nicht eine an- dere? Ferner, warum dieses Ding jetzt und ein andermal gerade nicht, und was früher langweilte, plötzlich gar nicht mehr? Es muß doch an all dem etwas sein, was uns langweilt. Was ist es? Woher kommt es? Was uns langweilt, sagen wir, verursacht Langeweile. Was ist dieses Verursachen? (For what does it mean that certain things and people cause boredom in us? Why precisely these things and that person, this region and not another? Moreover, why this thing now and not some other time and what about what once used to be boring and is no anymore. There has to be something in all of this that bores us. What is that? Where does it come from? That which bores us, we say, causes boredom. What is this occasioning?)".[46]

To have an application of the category of causality, conditions must be met so that when A happens, B happens. Now, this is precisely what does not happen. The fact that we are bored with the contents A, B and C or X, Y and Z does not mean that every time A, B, C, X, Y, Z occur we feel bored. It may happen only once. It may be episodic, sporadic. It can also happen that what did not previously cause boredom, now becomes deadly boring. The opposite can happen. A person annoys you and then ceases to be annoying. We do not really know what is it in these objects that cause us to be bored when they provoke boredom and

[45] Heidegger, *GA29/30:* 125: "Ist das so ein entsprechender Vorgang, wie wenn eintretende Kälte das Sinken der Quecksilbersäule im Thermometer verursacht? Ursache – Wirkung! Herrlich! Ist das etwa ein Vorgang, wie wenn eine Billardkugel an die andere stößt und dadurch die Bewegung der zweiten verursacht?"

[46] *Ibid*.

then disappear to no longer cause boredom. This set of problems makes it impossible to draw a causal link between the occurrences of A to Z and tedium. We can positively say that when one feels bored, a relation is established between A, B, C, or X, Y, and Z. The boredom is a mood or an atmosphere that we associate with an object, a person [14]. No question about it. The thing about this experience is the identification with its temporality: it drags on and on, time seems not to elapse because it is dry, it is empty and leaves us feeling empty. Any dull object has this relation with time and with emptiness: it is a retardation of life, it forces us to be dragged through time, and it is empty, it does not fulfil, it does not interest, etc. "Langweilig – wir meinen damit: schleppend, öd; es regt nicht an und regt nicht auf, es gibt nichts her, hat uns nichts zu sagen, geht uns nichts an." (Boring – we mean by this: it drags on, it is boring; it does not stimulate and stir, it provides nothing, it has nothing to tell us, it does not concern us.)".[47]

The boring drags us along an wears us out. In another formulation: it stops time and is draining, but these are still characteristics that we sense in us. They concern us. They alter us. Are they subjective? The interpretation of the subjectivity of these characteristics does not cancel the objectivity of their existence in things and in people that are boring when they annoy us. Subjectivity is not understood as something inner, impermeable, absolutely shielded. There is an interplay between mental reality and objective reality just as there is a personal interaction that admits dispositional transformations between people. We are affected by others as others are affected by us. This a priori atmosphere priori is already open as such and is as much objective as it is subjective, or neither objective nor subjective. Any exclusive characterisation of this emotional phenomenon would clearly not take into account what happens here. Just as we do not have a representation of boredom that is projected onto content to render it boring, we also do not have an actual imported boring thing that makes us feel bored. The same thing happens when someone annoys us or is always annoying to us or we annoy somebody or are boring to this person. There is not an interplay of boredom representations between us and others. We

[47] Heidegger, *GA29/30:* 126·

instantly feel boredom, that we are being bored or being boring. The emotion is already in the air if you will. Boredom is a possibility and that is the reason why time is experienced unfolding with content, but that can change. It can stop, it can be held back, it can be delayed. It can be emptied of its meaning. This interplay between person and thing, or between persons, is always just downstream of the opening to the existential horizon in which we live with emotions, affections, commotions, mental states, moods and dispositions with their cadences, vibrations, rhythms, and times.

It is important to stress the phenomenological origin of the analysis again. The progress of the analysis is slow but steady here. We are dispelling misunderstandings. The cause-effect relationship, the internalising of the mood as a mental phenomenon, and the metaphor are all operators that multiply the problems, when it is about the possibility of tuning into the mood as it is happening, i.e., trying to understand what it is telling us about the reality, our own selves, our relation to what happens to us in the emotional milieu. So, the starting point is the dispositional vibration that is felt, the affect, the emotion, the pathos, or whatever we may call it:

> "Wir sagen: aus einer Stimmung, aber nicht einer verursachten Wirkung; aus einer möglichen, uns möglicherweise befallenden Stimmung. Aus einer Stimmung her finden wir etwas so und so und sprechen es so an. Das heißt nicht: eine Wirkung und ihren Charakter auf die bewirkende Ursache übertragen. (We are saying: from an emotional cadence. Not, from an effect caused; a possible disposition that could possibly happen to us. From an emotional dispositional cadence, we find that something is like this and we talk about it. This does not mean: transfer an effect and its features to the effector cause.)".[48]

[48] Heidegger, *GA29/30:* 131.

7. Metaphors

Heidegger's third characteristic, which he refines, is the metaphorical. When we talk about dispositions are we speaking about things or metaphors of things? It is the same thing when we use allegories and parables. Words can either denote or connote. We know that, but experiencing that ground of the connoting or metaphoric usage of a word is ineradicable. The literal meaning of a word is one thing; the figurative use of a word is something else entirely. Yet when we speak of a dull person or a dull book, what are we saying? Is it not true that we are exporting mind phenomena and contents of a dispositional nature to persons and things? Further. The impressions that things and persons cause on us and the state they leave us in is then transferred to things and persons. There is a transmission chain that allows for dispositional reception of contents, emotional clipping of them, and then, at the appropriate time, we use that content to dispositionally describe what appears.

> "Langweilig, heiter, traurig (Ereignis), lustig (Spiel) – diese stimmungsmäßigen Eigenschaften, sie sind im besonderen Sinne subjektbezogen; nicht nur das, sie stammen direkt aus dem Subjekt und seinen Zuständen. Stimmungen, die die Dinge in uns verursachen, übertragen wir hinterher auf die Dinge selbst. (The attributes "boring, serene, sad, funny" are emotional, they exist in a relation to the subjects in a particular sense. Not only this. They directly originate something in the subject and its states. The dispositional qualities that things cause in us are transferred by us to the things themselves.)".[49]

Heidegger does not criticise metaphor as an expression of emotion. He is rather trying to show how the fundamental element of the understanding of meaning is always already implied in the metaphor. Therefore, if there is an excess of meaning expressed through the use of metaphor, it is not

[49] Heidegger, *GA29/30:* 127

the figurative meaning that comes after the literal, but the inverse. The figurative sense is primary in the very constitution and expression of meaning. It is not a simple matter. What Heidegger seems to be trying to say is that metaphor in the strict sense already presupposes the use of language, which is in its original sense metaphorical, allegorical, a parable. In Portuguese, the word for "word" is a translation of "parable". The literal corresponds only to a suspension of the second-order use of the metaphor which prevents the "figurative" use in the strict sense. Is scientific language, even mathematics, capable of a literal use of language? Does not what is equal, for example, or reflexive or transitive presuppose a sense other than the literal based on a world in which no two things can be presented as equal? Meaning and its expression are a priori in relation to which there can be literal and figurative senses, denotation and connotation. One does not start from a literal sense to a figurative sense, because the literal sense itself presupposes access based on the understanding of the sense itself. That is, what can be understood as a literal understanding is still metaphorical, an angle from which reality is seen, but which is not without equivocity. Now, the relation between meaning and reference can lead, on the one hand, to the neutralisation of meaning in order to stay only with the referent. After all, a = b if "a" and "b" have the same referent. "The evening star" is "the morning star" if, and only if, they have the same referent because "afternoon" and "morning" are completely different senses. On the other hand, a triangle is the same as a trilateral, although an angle is different from a side and only the 'tri-' makes it possible to understand that they are the same object. The fact of signification is once again the a priori, just as the figurative was first in relation to the literal.[50]

A smiling meadow, a serene room, and a melancholic landscape allow the disasso- ciation of objects from geography, architecture, and painting to be addressed according to the moods they may arouse in us at any given time. The cut-out smiling, serenity, and melancholy can mutatis mutandis be applicable to other objects according to the manner and mode of being of these same objects.[51]

[50] Cf. Gottlob (1892).

[51] But there is an aspect that should be mentioned, even if it cannot be developed here. The "matter" – and indeed the "form" of which emotions are made – is

8. Second Order

Analysis indicates a possibility that might pass unnoticed were it not to be taken up explicitly in the following paragraphs. It may well be that an emotional mood is there constituting both the reality of something or somebody boring and the subjectivity of the condition in which we each find ourselves. Yet, it may well happen that we do not realise its presence. "[es] ist sehr wohl möglich, daß wir uns beim Lesen gar nicht gelangweilt haben, nicht das "Gefühl hatten", daß in uns Langeweile bewirkt werde. ([it is] very possible that we were not being bored while reading, did not "feel" boredom being induced into us.)".[52] I may read a boring book. Yet, to read it, I have to somehow be pulled into the reading, be concentrating and understanding the main thread. A dull book does not necessarily prevent one from reading it. The same thing happens when watching a movie, a play, when visiting a museum, going to the beach, walking around, or meeting somebody. We might not realise that each one of these contents is boring. We have the experience of time dragging on, appearing to stop, and being delayed. On the other hand, it does not fulfil us, it is not full. It is, instead, an emptiness. We have dealt all our lives with tedium, that time that drags on and on to the point that it seems to stand still and to be a delay of life. We understand how it feels to cope with content which evacuates any sort of filling: "Aus einer Stimmung, von der wir dabei wissen, daß sie jederzeit aufsteigen könnte, die wir aber niederhalten, nicht aufkommen lassen wollen. (It is starting from the dispositional cadence, which we know could arise at any moment, but that we want to repress, that we do not want to let emerge.)".[53]

Yet, we unmistakably know tht something was boring back then. It was boring to read that book, to watch that spectacle; entire quarters

musical. Music as an acoustic expression captured by acoustic perception aims at a "tonic" event that we access directly as humans, even though we may not recognise it. It is not even necessary to be a musician to have this perception of reality. Music depends on time in order to be, just like any acoustic object. But music implies sound volume, sound quality, cadence, speed, rhythm. Life is temporal and acoustic.

[52] Heidegger, *GA29/30:* 130.

[53] Heidegger, *GA29/30:* 131.

of a city, or an apartment, or the décor of a home, but also times of the day, or days, or phases of life, or epochs were boring. The thing common to all boring things is that they drag on, they stop and drag us with them, they bring us to a standstill, they retard our life. Moreover, they are dull, they are boring, they do not stimulate, they have no interest. "Schleppend besagt: es fesselt nicht; wir sind hingegeben, aber nicht hingenommen, sondern eben nur hingehalten. Öde besagt: es füllt uns nicht aus, wir sind leer gelassen. (Dragged means: it does not hold our attention; we give ourselves away, but we are not taken, we are kept on hold. Desert means: it does not fulfil us, we are left empty)".[54] That which is drawn in comes to be translated ontologically as what delays (das Hinhaltende) and what is depicted as desert: what empties (Leerlassende). How we find ourselves, and the dispositional openness to this cadence, enables us to identify content as boring and the state we find ourselves in as boredom: "wie wir so und so angegangen wurden und uns dabei so und so befinden. (the way we were affected and how we find ourselves affected)".[55]

Thus, the longer or shorter duration of an emotion cannot be the criterion for de- ciding on its importance. It may happen that we have no other way of accessing the deeper dimension of emotions than when they surface consciously within a period. "Per- haps that boredom that so often slips by us is more important than the one we strive to annihilate" (Vielleicht ist gerade jene Langeweile, die oft nur gleichsam an uns verbei- huscht, wesentlicher als die, mit der wir uns gerade ausdrücklich abmühen)".[56] The fact that "it leaves us in an unpleasant and uncomfortable situation (in ein Unbehagen versetzt)"[57] may not mean anything. Neither duration nor emotional violence are necessarily criteria of truth. "Perhaps that boredom is more essential when it does not make us feel good or bad, but rather leaves us as if we were not even under any emo- tional presence (Vielleicht ist jene Langeweile wesentlicher, die uns weder gut stimmt noch mißstimmt und doch stimmt,

[54] Heidegger, *GA29/30:* 130.

[55] *Ibid.*

[56] Heidegger, *GA29/30:* 130.

[57] *Ibid.*

aber so, als seien wir überhaupt nicht gestimmt.)".[58] "This superficial boredom must lead us to profound boredom, or, to put it more properly, superficial boredom must reveal itself as deep boredom, by tuning us to the depths of existence. This passing, casual, non-essential boredom must become essential. (Diese oberflächige Langeweile soll uns gar in die tiefe Langeweile bringen, bzw., angemessener gesprochen, die oberflächige soll sich als die tiefe Langeweile offenbaren, uns im Grunde des Daseins durchstimmen. Diese flüchtige, beiläufige, unwesentliche Langeweile soll wesentlich werden.)".[59]

The relationship between surface boredom and time may not be immediately identified. However, when we get a brief glimpse of the phenomenon, we immediately understand this relationship. In these situations, we try to kill time, occupy ourselves with tasks. We feel that time is slow to pass. This lets us understand we get a feeling of time or, rather, time makes itself felt. Time "tells us" about itself by lengthening or shortening its pace, producing in us the sensation that it passes quickly or slowly. However, the time of boredom is different from the time when we cross episodes of boredom. The time of boredom is the time of existence itself. It never ends as long as we are alive. Surface boredom comes and goes. The time of authentic boredom comes from the depths of existence. What if the time of our lives was but an occupation of the free time of existence?

The relationship between surface and depth is very clear here. By putting the surface in relation to the background, we can understand that it is from the depths of existence that the condition for the possibility of experiencing surface boredom is constituted. The episodes of boredom depend upon deep boredom. Depth here means the a priori and transcendental condition of the being of boredom. It may happen that the ontic truth of boredom is circumscribed and isolated without allowing us to understand its profound dimension. The usual criteria must be put under this perspective. What is fleeting, constantly flying away, on the surface, is so determined in contrast to the background, which is the entire time of existence. Superficial emotions can indicate their depth.

[58] *Ibid*.

[59] Heidegger, *GA28/30:* 123.

The time from which they come is the time of life that we carry along the way. This time already existed when we came to life. Perhaps, then, everything is "inside" this time, which has existed at every moment of our lives. To understand what this time has to say to us, we need not resist it as soon as it emerges (Nichtalsogleich-Widerstehen) but let the emotions and feelings that accompany this time vibrate in its authentic cadence (Ausschwingen lassen.)

The emotional a priori allows s to find ourselves. The phenomenon indicates an openness that is different from the reflective one or the one given by self-perception. The experience of feeling emotions opens oneself to others[60] the world, oneself[61]. Not only that: time is the source of all our emotions, feelings, and moods. Is time an emotion? How can we have a perception of time if not through feeling it, through a sensation of time?[62] The causal explanation fails, because sometimes it works, sometimes it does not. There are many situations in which we do not identify any emotion at play. Looking back, though, we can have an insight into the shape of the emotion that was at work then. This second-order emotional dimension must be active for us to act the way we do. All our attitudes and behaviours are expressions of the emotional dimension that does not

[60] "The mobility and malleability of emotion are stressed by affect theorist Sara Ahmed. In the *Cultural Politics of Emotion* (Ahmed 2004a; 2004b), Ahmed argues that emotion is not a thing that originates or inheres within a subject, although we often speak of it that way. Nor does it inhere within an object, waiting to be released upon contact. Instead, Ahmed understands emotion as a set of relations between subject and object that defines both. Inherently fluid and shaped by power, emotions are not psychological states for her but instead social practices. Thus, she proposes, the question we should ask is not what affect is but what it does: how does it circulate within a society through its circulation? What sort of relations shape it or are shaped by it?" (Wohl, 2017). The analysis is obviously about a 5th century BC author. Emotions in antiquity were not just outside the subject, they were transcendence, connection to others when others appeared not as stray entities but were interpreted in the light of what they represent for us. The one we love lived for a time without existing for us as a possibility that did not even have an identity. In being loved a woman emerges as a goddess, as Aphrodite, an enemy emerges not only as a man but as the devil. Our fears and our loves do not arise only from our head but are collective entities like the boogey-man. Only modernity has confined emotions to the flow of the cogito and finds it extremely difficult to break the bonds of neuroscience.

[61] Is it not emotionally that we get estranged from the world? Cf. Hervy (2014).

[62] On the relation time-emotional life, cf. Florival (1987) and Lennon (2010).

show up but is active in a clandestine way. We know only too late what kind of emotional dimension (of feelings, passions, affects, moods) was active and working us out at any given moment in time. Somewhere, sometime, next year or in years to come, we will experience the constellation of emotions, moods, and feelings that are constituting our lives in this present moment. How can we actively get there, instead of just passively noticing this fact. What does this mean? Does the emotional dimension of our lives come from another world, from the future, is it teleological? How does it hurl itself on us? Does this dimension hoover above us to let itself be discovered? How?

9. Emotional Depth

There are no emotions that are exclusive to the past, emotions that are exclusive to the present, and emotions that are exclusive to the future. Emotions are not, as we have seen, just reactions to actions with corresponding responses. There is anticipation in emotions. They are pro-active, they open perspectives for the future. Without emotional perspective, the future is also cancelled. There are, without a doubt, emotions that give or seem to give more importance to the past: nostalgia for bygone times. In the present, we feel clearly cut stimuli: tension provoked by appetites of all kinds: hunger, thirst, addictive contents, but also sexual tension, irritation, and fury. In the present, we are exposed to all kinds of emotional stimuli. Even the memory of an episode from the recent past can disturb us strongly. We are exposed to all kinds of emotional provocations. Finally, hope and despair are emotions clearly grounded in the future: the excitement caused by the moment of anticipation, the promise of pleasure, and the threat that one feels coming from imminent danger.

Still, it is possible to understand the temporal shifting character of emotions not only as phenomena that take place in the past, present and future, but as events that take place with the past, the present and the future. On the other hand, the duration of an emotional episode can never be circumscribed by a chronometer. Although there are chronic depressions, there are also more or less sad and joyful phases in our lives that last for days, weeks, months and even years. Difficult mourning can last much longer, but there are also emotional phenomena that correspond to an epiphany, they allow for a turnaround in what seemed to be the set meaning of the course of a lifetime. This is how we find love or break up effectively with someone forever. What happens in an hour can affect our lives forever.

The radical origin of the deep emotional level is the future as possibility. We all have great hopes for our lives. We live with great expectations. This is the level of depth that gives rise to the apparently anonymous and subconscious nature of emotions. Trying to make them reveal themselves is the very work of the philosophy of emotions, and perhaps of philosophy as such.

A fundamental aspect in which I distance myself from Heidegger is his dogmatic use of the thesis of the three emotional levels: first-order or superficial, second-order or more profuse, and third order or depth. To begin with, it is not evident that there are only three orders and that the combinations are not more unclear and ambiguous. Moreover, one does not perceive the exclusive focus of the analysis to be on boredom. There would be the possibility of irradiating to other dispositions, or at least of referring the surface-depth structure to other dispositions. Heidegger refers to boredom as a fundamental emotion, but not as the only one. *Being and Time* lists guilt, bad conscience, death, fear and anguish as fundamental dispositions. However, the point to underline is this: from its deepest to its most superficial level the emotional plane reveals time in its happening. Time is emotion (which seems a little more bizarre and difficult to understand). While at the first level we can apply causal relationship to perceive boredom, or any superficial sensation, as a response and reaction to what happens to us, the same is not true of a second and third- order emotional response. Moreover, while in a second-order emotion I do not perceive the emotional structure that I later come to realise as the meaning of the situation at the moment I was experiencing it, a deep or third-order emotion sheds light on the future temporal element as preponderant. Still, there is multiple combinations of emotional levels that we experience in a single day, without realising why we go through these emotional moments. We have the perception that this is so. So, we try to understand why we feel the way we do. It is precisely when it is difficult to understand the meaning of our emotional situation, that is, when we do not understand why we find ourselves in it, that we try to find the key to that understanding. We can have a picture of this in an episode of William James' situation of unrest described by Scheler.[63]

On certain weekday afternoons, James was required to give a course on formal logic. Logic, especially formal logic, was not even remotely an interest of William James'. Now, one particular morning, he noticed quite early on that he was feeling especially moody, "jumpy", with an eagerness to go about doing this and that without being able to concentrate on a specific task. He was, as it were, experiencing a sort of self-prescribed occupational therapy. James paced back and forth in his office and all around the house, collecting bits of paper, sharpening pencils, sitting down and getting up, and so on. Then, he tried to figure out the reason for his state and wondered if it could be due to having read late into the night. He knew perfectly well that he was going to teach the class that day in the afternoon, but it did not even occur to him that the state he was in was the result of an appointment, of a class scheduled for a certain time, a future time relative to the early morning of that same day. Where does the psychological "influence" of emotional states we often find ourselves in come from if not from the future? Are there not countless examples of similar situations that we all experience? Scheler examines the case further. It is not just an isolated instance of something we do not want to have on the agenda that makes a bad impression on us and leaves us in a bad state. It is the whole indeterminate future that is creating pressure on us, but we cannot then make a mathematical induction: if it happens that I am in a bad emotional state because of a future event this means that all future events cause a precarious emotional state. What happens is the opposite. It is because there is a future that all time can already constitute itself now, just as a moment ago, before that, yesterday, the day before yesterday. The future as possibility, beyond any limit, is ever since emotionally and anonymously structuring my whole life. Most of the time, and primordially, I count on "my future" without any reflexion upon it. However, when experiencing

[63] Cf. Max (1913).

despair and anxiety, or dull emptiness, the whole time of the future is cancelled and its cancellation already makes its effect felt now. We often say that one person has a future and another has no future at all, and we know how good prospects bring joy and brightness, thus colouring all our experiences, and how bad prospects cloud everything. We do not need to have any clear representation of what is going to happen, of what our perspective on reality is. However, the fact is this: we are subject to a perspective that opens up from the future; there is a retrospective that comes from our future and that makes our whole future, our whole present and its content, our whole past and its content, our whole life and its content bright or sombre. So, there is also a prospect already at work, even if we do not realise it, opening up to a future moment in time, even if we do not live to see it happening. A recalcitrant experience allows us to perceive the future of an emotion. In surviving trauma, we understand that everything is going to be different from then on. When old age strikes, when we understand an episode that makes everything irreversible, we get a glimpse of the never-again, we understand what forever means.

Or rather, I now have all the ime in the world, all the time in the universe, all the time of eternity. Yet I cannot make anything of it. What does it mean not to be able to make anything of time? It means not knowing how to fill it or else how to kill it, how to make myself not feel unoccupied for all the time of my life. What deep boredom indicates when it appears is the same as the anguish of death. Boredom empties and paralyses, it brings emptiness and inanity with it. It interrupts my life abruptly and totally. It is that definitive, empty feeling that envelops all my emotions, all the moments across which my life is distributed. Boredom exerts pressure on us and manifests itself in such a way that we know we are having a bad time. Deep boredom turns every day of our lives into a Sunday afternoon with nothing to do. Our whole life is that Sunday afternoon. We have all the time in the world and do not know how to enjoy a single moment of it. Is there no emotion that can rescue us from this tedious moment? Is the emptiness of the future the dimension that runs parallel to all the instants of our life with or without felt emotions? There is not a moment in our lives that is not founded on this emotional depth that overflows outside the boundaries of my life and reaches into others' past and future lives. Permeating everything with its powerful "no", this sense of emptiness that comes from being bored makes us ask ourselves the question of meaning. How is it with you? Have you been living a good life? Has it been worthwhile?

The sub-conscious emotional level is detected in the present moment when one feels discomfort, malaise, restlessness. However, the emotional provocation comes from the future. Now, how is it that a felt emotional phenomenon is the effect of a cause that lies in the future. Is this not an inversion of the natural understanding of causality? Is not the cause in the past? Is not the effect of a cause its consequence and its future? Emotional depth comes from the future. The emotional foundation lies in the future. The emotional situation we are in comes from the future. It is in the future that we have to look for the reason why we are like this. The answer can be found in the fact that in (just) a moment we are going to go through a situation that is already stressing us. Yet we are not having any representation of that future scene. Without a thematization or representation of what is going to happen, or of how the future is going to happen, we are already emotionally metamorphosed. What is harder to understand is that, as a future, an emotion is a mere possibility. Yet it is a possibility that can be far more effective than any reality.

We feel the pressure of the hour when we will be doing something we do not want. We know quite well what we have scheduled. It is different from the indeterminacy of the future of a Thursday at 5 p.m. for which we have not scheduled anything, or from my present discomfort at the idea of teaching 6 p.m.–9 p.m. classes in September when it is still August. We are thus always already under the pressure of a future moment. For Heidegger, the fundamental question is that somehow deep boredom transforms life. "Sunday afternoon in a big city" is an expression of emptiness, of inanity, of the total suspension of time in my life.[64] When something like this happens, it is not only my inner mental psychology that is transformed, but the whole "world", the whole "universe", everybody else, everything else. Life itself shows up and reveals itself. It

[64] "One feels it is boring, when on a Sunday afternoon one goes for a walk through the streets of a big city. (es ist einem langweilig, wenn man an einem Sonntagnachmittag durch die Straßen einer Großstadt geht.)" (Heidegger *GA 29/30*, 204)

is a moment of revelation, a moment of being. Maybe now we can ask "what is the meaning of being?" or is it the other way around? Is it not that Being Itself asks us: what is up? What is going on? What are you up to in your life? One needs to get close to those phaenomena. Sometimes deep emotions surface in our conscious life. It is a matter of fact, but the work of philosophy is to make them show up and to allow us to live in such a dimension where being unleashes itself emotionally.

References

Ahmed, S. *The Cultural Politics of Emotion*; Routledge, Taylor & Francis Group: London, UK, 2015.

Anders, G.S. Emotion and Reality. *Philos. Phenomenol. Res.* 1950, *10*, 553–562.

Aristotle. *The Metaphysics*; Aristotle in 23 Volumes, Volumes 17, 18; Tredennick, H., Ed.; Harvard University Press: Cambridge, MA, USA; William Heinemann Ltd.: London, UK, 1989.

Aristotle. *The Nicomachean Ethics*; Aristotle in 23 Volumes, Volume 19; Rackham, H., Ed.; Harvard University Press: Cambridge, MA, USA; William Heinemann Ltd.: London, UK, 1934.

Bruss, K. Searching for Boredom in Ancient Greek Rhetoric: Clues in Isocrates. *Philos. Rhetor.* 2012, *45*, 312–334.

Capobianco, R. Heidegger and Jung: Dwelling Near the Source. *Rev. Existent. Psychol. Psychiatry* 1993, *21*, 50–59.

Davis, W.A. Expression of Emotion. *Am. Philos. Q.* 1988, *25*, 279–291. Available online: http://www.jstor.org/stable/20014251 (accessed on 22 July 2022).

De Lauri, A. Boredom and Crisis in the Humanitarian Realm. *Anthropol. Today* 2014, *30*, 23–25.

Elpidorou, A. Moods and Appraisals: How the Phenomenology and Science of Emotions Can Come Together. *Hum. Stud.* 2013, *36*, 565–591.

Florival, G. Vie affective et temporalité. *Rev. Philos. Louvain* 1987, *85*, 198–225.

Gilje, N. Moods and Emotions: Some Philosophical Reflections on the 'Affective Turn'. In *Sensitive Objects: Affect and Material Culture*; Frykman, J., Frykman, M.P., Eds.; Kriterium: Gothenburg, Sweden, 2016; pp. 31–53.

Gottlob, F. Über Sinn und Bedeutung. In *Zeitschrift für Philosophie und Philosophische Kritik*; Pfeffer Verlag: Leipzig, Germany, 1892.

Hatlen, B. Oppen and the Unspeakable. *Paideuma* 2013, *40*, 211–253.

Heidegger, M. *Prolegomena zur Geschichte des Zeitbegriffs (History of the Concept of Time: Prolegomena)*; Heidegger, M., Kisiel, T.J., Eds.; Indiana University Press: Bloomington, IN, USA, 1994.

Heidegger, M. *The Fundamental Concepts of Metaphysics: World, Finitude, Solitude*; Heidegger, M., Ed.; Indiana University Press: Bloomington, IN, USA, 2012.

Heidegger, M.; von Herrmann, F.-W. *Die Grundbegriffe der Metaphysik Welt – Endlichkeit –Einsamkeit*; Vittorio Klostermann: Frankfurt am Main, Germany, 1983.

Hervy, A. Émotion et aliénation. *Études Sartriennes* 2014, *17*, 21–40.

Jensen, K.A.; Wallace, M.L. Introduction: Facing Emotions. *PMLA* 2015, *130*, 1249–1268.

Lennon, K. Re-enchanting the World: The Role of Imagination in Perception. *Philosophy* 2010, *85*, 375–389.

Malabou, C. How Is Subjectivity Undergoing Deconstruction Today? Philosophy, Auto-Hetero Affection, and Neurobiological Emotion. *Qui Parle* 2009, *17*, 111–122.

Martha, F.; Miller, M. Object Emotions. *Symplokē* 2016, *24*, 155–170.

Max, S. Zur Psychologie der sogenannten Rentenhysterie und der rechte Kampf gegen das Übel. In *Vom Umsturz der Werte. Abhandlungen und Aufsätze GW Gesammelte Werke III 2007*, 1st ed.; Francke Verlag: München, Germany, 1913; pp. 236–238.

Mendonça, D. What a difference depth makes. *Rev. Filos. Aurora* 2019, *31*, 671–694.

Purton, V. Tennyson, Heidegger, and the Problematics of "Home". *Vic. Poet.* 2012, *50*, 227–248.

Sheets-Johnstone, M. The Enigma of Being-Toward-Death. *J. Specul. Philos.* 2015, *29*, 547–576.

Solomon, R.C. On Emotions as Judgments. *Am. Philos. Q.* 1988, *25*, 183–191.

Stevens, W.; Serio, J.N.; Beyers, C. *The Collected Poems of Wallace Stevens. Vintage Books*; Alfred A. Knopf, Inc: New York, NY, USA, 2015.

Woermann, N.; Rokka, J. Timeflow: How Consumption Practices Shape Consumers' Temporal Experiences. *J. Consum. Res.* 2015, *41*, 1486–1508.

Wohl, V. Thucydides on Political Passions. In *The Oxford Handbook of Thycydides*; Balot, R., Forsdyke, S., Foster, E., Eds.; Oxford University Press: Oxford, UK, 2017.

Situating Mental Depth

Robert W. Clowes and Gloria Andrada

Clowes, R. W. & Andrada, G. (2022). Situating mental depth.
***Avant*, *13*(1), 1-30.**
https://doi.org/10.26913/ava202203

1. Depth psychology and its critics

Hilary Mantel´s Thomas Cromwell is a capacious literary creation. An inhabitant of Tudor England who can comport himself within any social echelon; Cromwell contains multitudes that are all himself. He is the butcher's boy, the soldier, the hired tough, the accountant, the lawyer, the parliamentarian, the diplomat, and the confidant and advisor to Kings – the architect of the English Reformation. Cromwell is an adept social practitioner and manipulator, always in action, an anti-Hamlet. Hamlet thinks, dithers (and quips), dissects himself but cannot act, and in so doing gives us a Western literary and moral archetype of a certain kind of inwardness of mind. Mantel's Cromwell thinks, reflects, quips, and engages in the minds of others. He regrets but finds resolutions, he shapes events, he baulks at simplistic revenge, and relentlessly, he acts some more; always planning, plotting, and shaping the landscape of Tudor England to himself. Cromwell is also archetypal – though an unusual version – of the notion of the "rounded" literary figure that depicts something of the contemporary self-understanding of what constitutes a rich mental life. His depth can be taken as a depiction and echo of the complexity and richness of the life of mind, which we all possess and of which we are so intimately acquainted. He captures something about what we – perhaps at our best – take ourselves to be. We, modern human beings, assume ourselves to be creatures of profound inward *mental depth*.

The idea of mental and psychological depth can be traced to intellectual trends of the late 19th century and perhaps especially to the work of Sigmund Freud. The Freudian idea of a structured 'dynamic' cognitive economy involving the interplay of conscious, semi-conscious and unconscious parts has given us foundational elements of much contemporary folk-psychology. It continues to shape both the 'manifest image' of what it is to have and be a human mind and also lies behind many ideas in contemporary psychology and cognitive science.[1] This

[1] Although Freud is now taken as the main exponent of this idea, other contemporary or near contemporary figures such as William James, Carl Jung, Eugene Bleuler, and Pierre Janet all posited different versions of "depth

should not come as a surprise. Folk-psychology is in a constant interaction with scientific psychology and when new explanatory notions emerge, such as the notion of the unconscious, we frequently incorporate them in our folk explanations and, consequently, they become part of the tools we use for our individual and collective self-understanding.[2]

Even if much of the psychodynamic framework bequeathed by Freud has fallen from favour, the idea of the deep and unconscious background to many of our psychological processes is still undeniably influential in much cognitivist theorizing about the mind.[3] In fact, there is a widely regarded historical view that Freud´s idea of the unconscious was somewhat reinterpreted by cognitive psychology (see for instance Power and Brewin, 1991, and Westen, 1996),[4] where there is a widespread commitment to the view that cognitive processing takes place unconsciously or sub-personally.[5]

psychology" where the sources of motivation, ideas, and the self could be traced to the operation of hidden, unconscious, or (sometimes) subconscious forces of which the subject is unaware.

[2] The notion of the unconscious deeply influenced folk-psychology in the late 19th and early 20th Century both through general cultural discussion but also through works of art from the paintings and films of Surrealists such as Salvador Dali, Rene Magritte, Luis Buñuel, and to the work of vastly popular film-makers such as Alfred Hitchcock. It is difficult to describe much of the cultural life of the first half of the 20th Century without reference to the notion of the unconscious. But the ideas also became part of folk-psychology. Many people who came into contact, even quite indirectly with these cultural trends reinterpreted their own mental life in the light of the new ideas of depth psychology.

[3] The development of psychoanalysis in its various schools, but especially those influenced by the central figure of Freud, was based on the core idea of bringing the unconscious into the light in order promote cognitive change in individuals struggling to understand themselves or change their behaviours. It is worth noting that, though drawing from a different perspective, the concept of depth also played an important role in phenomenology, and more particularly, on the influential work of Merleau-Ponty (1962). Merleau-Ponty explored the relation between depth and agency. He introduced the notion of 'primordial depth' as a basic form of human experience. Depth here is not something internal, but a constitutive element of being an embodied subject that is "involved in the world" (1962, 256–7). In section 4.2, we will come back to this, when we motivate our embodied approach to mental depth.

[4] Thanks to an anonymous reviewer, for encouraging us to clarify this further.

[5] For example, the computational model of mind in canonical forms (e.g., Fodor´s *Language of Thought hypothesis*) presupposed that the computations that take place in the mind are largely unconscious (see Fodor, 1975).

Summing up, on a depth conception, our conscious minds are framed as being only the tip of a cognitive iceberg, while the lion's share of our mental life, the sources of dreams, our creativity and eureka moments (but also our hidden desires and biases) take place behind and below the functioning of the conscious mind, in what is often described as "the vast ocean of the unconscious."

This vision of mental depth has come in for a sustained challenge in recent times. For some, the tip of the iceberg metaphor is badly mistaken. The central idea is that mental depth, at least as standardly conceived, may be much more appearance than reality: it is a kind of illusion, that we are only able to maintain through the mind's inbuilt blindness to many of its own gaps and absences.[6] Our mental life may in fact be thinner, less substantial, and much gappier than supposed by the intellectual giants of the nineteenth century, and literary novelists from Flaubert to Mantel. This view occurs in several places in the contemporary cognitive science literature (Dennett, 1991; Blakemore, 2002), but it is perhaps most clearly articulated in Nick Chater´s (2018) book *The Mind is Flat.* According to Chater, the kind of mental depth that Hillary Mantel's Cromwell suggests is a self-flattering but largely fictive depiction of the sorts of beings we are. According to Chater our own mental depth is just as illusory as the apparent depth of fictional creations.[7]

On Chater's account, we emerge as very different sorts of creatures from that imagined by the giants of 19th and early twentieth century psychology and literature. Rather than the sources of our cognitive prowess being hidden away, we are instead ceaseless improvisers. Our brains are always engaged in producing one pattern-completing Gestalt at a time. But, more worryingly, we are also ceaseless confabulators, forever making up fictive reasons and motivations for our actions but

[6] For some discussion of whether gaps and absences in the self really imply illusions about self, see Clowes & Gärtner, 2020.

[7] There are of course many authors who argue that the self is fictional, or a sort of illusion (Dennett, 1992; Hume, 1978 [originally 1739]; Metzinger, 2004). But the claim that mental depth is illusory is a distinct – if related – claim, to the claim that the self is an illusion. The idea of the illusion of mental depth goes beyond the nature of the self to the claim that the mind is itself a more gappy, less coherent entity with less knowledge about its own nature than we generally suppose.

blind to the gaps and absences which abound in our conscious mental lives. If Chater is right, the sort of deep inward minds we find in, e.g., Hillary Mantel's Cromwell are doubly fictional.[8] Much of what we take to be the depth of the human mind is itself an illusion or fiction.

Although we believe there is much to be admired in Chater's critique of mental depth, at least as standardly conceived, in this paper we will present the case against his view. We argue that there is a kernel of something correct in Chater's idea of the just-in-time, improvised character of cognition, yet this does not, in itself, add up to a critique of mental depth *per se*. Instead, we use Nick Chater's ideas as a springboard for creating a new understanding of mental depth.

Our account especially draws upon recent work that identifies the origin of many aspects of the human mind through its dependence upon dense patterns of skillful interaction within a rich artefactual and social environment. More precisely, we argue that the characteristic mental depth of the human mind emerges in the practice of skillful actions. It will take a little work to illustrate what we mean by this idea, so we ask the reader to bear with us as we gradually develop our account here. For now, we observe that the notion of depth can be rebuilt in ways that is scientifically progressive but that at the same time retains some elements that are deeply entangled in the image we have of ourselves as minded creatures. As we will show, our renewed notion of mental depth is not best described as an inner mental depth but a depth that is situated, skillful, and active. It is in our skillful interaction with a rich artefactual and social world that our mental depth unfolds.

Our plan is as follows. We begin (§2) by presenting in more detail the illusionist challenge to the traditional notion of mental depth put forward in Chater (2018). In the following section (§3), we develop a vignette regarding the acquisition and refinement of skilled practices designed to help us illustrate what we claim are the *real* sources of mental depth, which we argue is largely to be accounted for in terms of the depth of skilled situated practices. We examine

[8] Thomas Cromwell is a fictional creation in quite a complex sense. He was also a historical figure in the court of Henry VIII. When referring to Cromwell in the paper we are primarily referring to the fictional creation (Mantel, 2011).

one concrete form of the acquisition of mental depth in a particular domain through learning to play the cello with the Suzuki method. The next section is devoted to our own alternative approach to mental depth (§4). Essentially, we argue that mental depth is real, and we can find its sources in two places: the depth of hierarchical predictive knowledge (§4.1), and the depth our embodied skills and the situations in which we are embedded (situated and embodied depth) (§4.2). In the final section (§5) we return to our proposal in order to explain why mental depth in its practiced reality should not be understood as a confabulation.

2. Illusionism and the flat mind

Our account of mental depth is developed against the particular backdrop of *illusionism* (Chater, 2018; Frankish, 2016) and more precisely the form of illusionism developed by Nick Chater in his book *The Mind is Flat*. The cornerstone of Chater´s argument is that the inner life as we imagine it, or at least as it is depicted in literary novels, but also in cognitive psychology and much folk psychology is largely illusory. A central target of Chater's book is therefore to debunk a certain conception of *depth psychology*. For Chater, the conscious mind is not the tip of the iceberg with the main edifice of our thought hidden away (see Chater, 2018, p. 186). Rather, for him our brains are Gestaltist parallel processing systems that produce just one coherent thought, perception, and interpretation of the world – as a conscious deliverance – at a time. The sense we might have of a dense background of thought behind this, which only occasionally percolates to the surface, is illusory. Rather our apparently deep thoughts are forged in the moment we encounter a rich and complex world in need of interpretation.

Chater's aim is to show us that the folk picture of mind is wrong. Our privileged access to a private inner world of self is much less firmly grounded than folk psychology regards it as being. Human consciousness is more gappy, improvised, and low bandwidth than we (the folk) suspect. Chater's argument is a form of illusionism, in that he claims that we are mistaken, and operating under an illusion about the nature of our minds and cognitive processes. In a sense, the depth of the human mind is just as much a fiction as the literary creation of Thomas Cromwell.[9]

On the face of it, such a view may seem radically at odds not just with folk-psychology, but also with much phenomenology. Let us look at these ideas and some of reasons for them in a more articulated way.

2.1. The case for the depth illusion and why Chater argues the mind is flat

Chater uses a series of examples to build the case that our sense of mental depth and even the apparent *coherence* of thought are illusory. Two examples here will suffice to make his case clear. One involves our sense of the reality of fictional creations. Chater's book – rather like our paper – begins with the extended discussion of a fictional creation, in his case Anna Karenina. Anna is the titular figure of the novel, and one of its central characters, and in the terms, we are using here a *deep* literary figure.[10] Chater points out that although the attentive reader of Tolstoy's novel will certainly feel themselves to have a deep acquaintance and understanding of Anna, it may then come as a surprise to learn that many characteristics – it must be said, rather superficial characteristics of Anna such as her hair color, her height, her build etc. – are never explicitly stated in the book. However, when on reflection we discover that we do not know these characteristics, and indeed that we never noticed that we do not know these characteristics, we are surprised.[11] This is an indication that although we think we know how Anna looks and many other aspects of her person, this is an illusion.

[9] The exact sort of illusion that Chater holds to be the case is close to that originally stated by Dennett in his book *Consciousness Explained* (Dennett, 1991) especially where Dennett's ideas shade into the idea of the Grand Illusion (O'Regan, 2002). As we read Chater, his views are a little more distant from some other contemporary forms of illusionism that hold either that there is no stream of consciousness (Blackmore, 2002), or that qualia is an illusion (Frankish, 2016).

[10] In nomenclature used by literary theory we could also say she is a "round character."

[11] For a discussion of Dennett's classic paper on surprise as philosophical methodology, see Dennett, 2001.

The second example draws upon another series of novels, in this case, Gormenghast (see Chater, 2018, p. 21). The Gormenghast novels are famed for, amongst other things, the detail of the descriptions of physical locations and the sense of reality they convey. However, Chater points out that although we may feel we have a deep and coherent sense of the novel's locations and how they fit together, the idea that we can have a fully coherent idea of this cannot be right, for the layout of the castle has been shown to be inconsistent (Chater, 2018, p. 21). The *feeling* of coherence, conveyed by the accumulation of surface detail, makes us think we coherently imagine the castle locale of the novel, but it is an illusion. Crucial for the argument, as it goes for our feelings of the depth and coherence of fictional worlds, so it goes for our feelings of depth and sense of our own inner lives. It may feel like we have a deep coherent inner life behind and below conscious experience but this too – so Chater claims – is illusory.

Chater also develops a series of examples designed to show that our sense of many of the objects of perception and knowledge are much more gappy, sparse, and inconsistent than we typically take them to be. Pride of place in many of these examples is that class of visual illusions that are engendered by impossible figures (some examples can be found in the visual illusions from M.C Escher, such as the famous "Relativity".). For Chater, the sorts of experiences engendered by impossible objects rely on an apparent coherence which is not really in the scene, or perhaps better in only local elements of the scene. The global scene is of course, by definition, *incoherent* – at least in the sense of confirming the earthly sense of spatial geometry and gravity. But on closer scrutiny, Chater argues, we find that there is no overall coherence and indeed our sense of depth is itself illusory.

The main idea is that impossible figures are not just mysterious visual games but suggest something profound about our visual systems and the nature of our minds more generally. Our minds tend to project a depth and reality which is not really there. They presage the mind´s tendency to mistake the overall coherence and depth of presentations of the world. In this vein, Chater writes:

> When viewing an impossible object, we have the overwhelming sense that we are looking at a 3D scene, albeit

> a peculiar one. But this 'feeling' of solidity is completely misguided – we are actually looking at a flat image that has no possible 3D interpretation. This is yet another illustration of the illusions of depth. These illusions of depth, which can be both literal, as with impossible figures, and metaphorical, as with stories and explanations, are everywhere. (Chater, 2018, p. 39).

The moral of this illustration is that the sense that we have of the background detail, of proximal depths of the contents of our minds: our perceptions, our thoughts, our memories, and indeed sense of self are all much more like a sense of depth of literary figure like Anna Karenina, or the sense of coherence of an M. C. Escher lithograph. It is apparent but not real. That is why for Chater, the ultimate illusion of depth is the depth of our own mental states, and the depth of our minds.

It is worth noting that these ideas echo previous discussions of gappiness or discontinuities of consciousness such as the idea of *The Grand Illusion* first framed around some puzzling phenomena first discovered in perceptual psychology focusing on phenomena such as change and inattentional blindness (e.g., Simons & Levin, 1997; Simons & Rensink, 2005), our experience of the perceptual world is much less high bandwidth and detailed than we take it to be. Marshalling a multitude of such studies, Chater claims that the human mind in general has much less of a grasp on detail than we think, but not just on the detail of perceptual experience but of the content of our minds.[12]

According to Chater, the illusion of mental depth is also manifested in the idea of *background processing*. This idea has its antecedents, on the one hand, in the Freudian idea of the unconscious which can be pictured as a background and quasi-personal stream of thought coming to its own inaccessible determinations and holding its own beliefs and desires separate from the conscious access of the ego. On the other, with the idea of background processing which is deeply implied in much traditional (and contemporary) cognitive science. Against this, Chater's mantra is "no background processing," the idea being that the mind is an

[12] In this respect, Chater concludes: "the mind itself is an impossible object" (p. 21).

improviser always making up the best possible, multi-modal and – so far as is possible – integrated interpretation of whatever it is currently being encountered. Instead of there being a constant active background, the brain produces such acts of interpretation one chunk at a time, thereby creating the illusion of a deep mental life.

A core example here is Kekulé's discovery of the structure of Benzine rings, the typical explanation of which Chater takes to be a myth.[13] After working on the problem of the structure of Benzine rings for many months Kekulé fell asleep gazing into the fire whereupon he had a vision of snakes amid the flames, one of which reached back upon itself and bit its own tail. With this vision Kekulé finds himself wide-awake, inspired and with the solution to the problem of the structure of Benzine alive before his mind´s eye. Benzine has, of course, a ring-like / hexagonal structure with each carbon atom bonded to two other carbon atoms and a single hydrogen atom (C6H6). Kekulé's discovery of this structure is taken as canonical evidence of the idea of background processing. While Kekulé´s conscious mind was resting (indeed asleep) his unconscious mind was said to be working away at the problem. For Chater, this is a sort of post-hoc confabulation of the creative process, and thus the idea of background processing is an illusion.

What then is cognition really like? Chater defends a "cycle of thought" analysis of conscious experience whereby the brain produces one conscious and apparently coherent impression of the sensory deliverances at a time.[14] According to the cycle of thought hypothesis, the mind is able to focus on just one overall interpretation of events and our worldly interactions at a time (we might say a single Gestalt!). The massively parallel connectionist architecture of our brain is constantly tasked with producing one global overall interpretation of the world at a time. All processing power is brought to bear on this, but this leaves nothing left over for independent thought processes going on in the background. A crucial aspect is that the brain is able to handle just one chunk of the world, one perceptual event, one thought at a time. This challenges some central assumptions of cognitivism, which holds that

[13] See Chapter 9: The Myth of Unconscious Thought.

[14] For a related early account of the cycle of thought see McCrone (1999).

there are lots of processing or "thinking" going on below the level of consciousness.[15] But there is no background processing: just one cycle of thought after another; one element of thought at a time.

The question then becomes: why then do we have this feeling of depth if we do not have it in reality? According to Chater's account, there are two reasons at work here. At one level, our brains are rapid and voluminous improvisors. This means whenever confronted by a particular scenario in need of explanation, our brains are poised and ready to fill in all of the gaps, giving rise to the illusion of mental depth. Much of what we take to be background processing is better explained by a sort of just-in-time filling-in. The other central part of this story is *confabulation.* While our brains are only producing one overall picture or Gestalt at a time, we are at any point able to turn this interpretational process back on ourselves in order to interpret what we must have felt, believed, or thought to achieve the cognitive processes we just did. However, the stories produced in such acts of auto-interpretation or self-explanation, are largely – or perhaps entirely – confabulated.

To summarize, mental life is not some sort of detailed internal picture, or a Cartesian Theatre (1991) as Dennett says. Our mental life is much sketchier and gappier than this (see Chater, 2018, p. 52). In

[15] In connection with how to relate conscious thoughts and the presumed mechanisms that produce them, Chater writes "There are no conscious thoughts and unconscious thoughts; and there are certainly no thoughts slipping in and out of consciousness. There is just one type of thought, and each such thought has two aspects: a conscious read-out, and unconscious processes generating the read-out." This formulation is Chater's attempt to circumvent the difficult question of how to relate unconscious processing and what we take to be the conscious contents of the mind at any moment. The problem with this way of expressing things is that it tempts the Dennettian question: who exactly is looking at the read-out? The idea of Cartesian Materialism was developed by Daniel Dennett (1991) precisely to serve as a warning to philosophers and cognitive scientists who assume that there must be a place, conscious experience, where it all comes together. Such a view can become misleading when there is a suggestion that there is an output screen or read out that expresses conscious thought to someone. Who exactly? An inner homunculus? In this paper we do not endorse any particular theory about how content becomes conscious in the brain or to what purpose, but we hold open the Dennettian possibility that there is no place where it all comes together. Perhaps consciousness is, as Dennett suggested, better understood as multiple drafts rather than a single canonical stream. For one way this could be cast into a predictive processing framework, see Dołęga & Dewhurst (2021).

addition, we are largely deceived about the coherence, detail, and even existence of much of our mental life. Chater writes at one point that "the unavoidable conclusion of these finding is that the mind itself is an impossible object" (Chater 2018, 51). Much of the detail of such mental imagery is filled in as and when needed, through more operation of the cycle of thought. Thinking and feeling on this analysis is a sort of improvised, perceptual, just-in-time sort of process. We can only hold one thing in our head at the same time. But we are able to confabulate a back-story for all of our cognitive episodes in line with whatever is our best interpretation and what folk-psychology says. We simply do not notice the many inconsistencies in our mental imagery or even in our visual perception, except where these are pointed out to us where we are shocked or surprised (see also Dennett 2001). We are constantly confabulating a story about our mental processes which is largely at odds with reality. The mind is flat, but we are great confabulators of depth.

2.2. The challenge of mental depth

Chater´s critique lays the groundwork for a novel and challenging view of the mind that seeks to rethink much of what we take to be the nature of the human mind and self. To conclude this section, we want to draw attention to the elements of his view with which we agree and those where we think he goes wrong. We want to read his view, against the grain, ultimately not as an elimination of human mental depth, but as a doorway through which we can see its real sources.

To clarify this, we largely agree with Chater's critique of the Freudian notion of depth and a unified cognitive unconscious. But as we will show, this is not the only notion of mental depth available. In a similar vein, we believe that Chater is largely correct to criticize standard cognitivist views on background processing. However, as will later become clear, mental depth need not be construed as internalist, fully private, or even classically cognitivist in the ways that Chater calls into question. Moreover, his more positive emphasis on the improvisational aspects of mind indicates at least some of the sources of our cognitive prowess. But we do not think that an inference to the lack of mental depth necessarily follows from these insights.

Consequently, our main critique of Chater is not that his picture of a constantly improvising mind is wrong, but that the way he understands mental depth undermines the alternative vision he wants to replace it with. In other words, while Chater's idea of the just-in-time, improvised character of cognition is sound, this does not, in itself, add up to a critique of mental depth *per se*. Depth is not the result of serial background non-conscious thinking processes, but its source lies in our rich and deep interpretative capabilities, the depth of the situation and especially the nature and structure of skillful action. Through the development of skills, we are able to attune ourselves to experiencing an expanding set of features of the world, both larger coherent wholes, and episodes of coherence spreading out over longer vistas of time (e.g., Donaldson, 1979, 1992). Essentially, the development of skillful practices reveals new and rich dimensions of the world as we encounter it. Such *mental depth* can be found in the novel and more sophisticated modes of interacting made available through the acquisition and refinement of skills. These practices, moreover, are not illusory nor confabulated. They are real and in the world, and emerge through the intertwined process of coming to rely on the rich affordances of the local environment; especially those made available by artefacts and the patterns of action we build around them.

Building on this, we propose an alternative embodied and situated approach to mental depth. We argue that this approach is immune from the sorts of critique that Chater has staged. Moreover, we stress that mental depth is real, and we can find its sources in two places. First, in our acquired abilities to *see* depth through our accretion of hierarchical predictive knowledge which simultaneously structures how we perceive and act on the world (the perceptual aspect of depth). Second, in the depth of the situations in which we are embedded and to which we learn to skillfully respond. To motivate our account, we will now move to discussing a concrete instance in order to analyze how and where mental depth shows up.

3. Finding depth in skillful practice

In this section, we aim to draw out some of the special aspects of the human mind that show up when it is examined in its ecological setting, namely through the dense patterns of skillful interaction with a rich artefactual and social environment. In particular, we want to illustrate the sorts of skillful situation-engaged minds we have and locate what makes human cognition special by examining the appropriation and refinement of skills. Human cognition takes place amid a set of densely integrated interactions orchestrated between human beings, their artefacts and each other, in a rich cultural environment. Such settings can make possible capabilities of the human mind which appear absent in the laboratory (Donald 2001). Often, we are not very conscious of the contribution that the artefacts make, but, at least within cognitive science circles this has started to change (Clark, 2008; Malafouris, 2013; Norman, 2000). Moreover, these interactions do not take place in a vacuum, but are guided – in ways we will shortly explore – by cultural practices (Hutchins, 1995, 2011; Menary, 2007, 2018).

In what follows, we introduce a vignette aimed to help us explore the situated nature of mind, that is, the interleaved interactions of an artefactual and social context in the development and refinement of cognitive depth in a socially and artefactually constrained context. This (phenomenologically informed) vignette is designed to help the reader understand what we mean by the situated depth of the human mind and form the basis of the way we will then seek to face the illusionist challenge more directly. To illustrate these ideas, we explore the concrete example of a child learning to play the cello guided by the Suzuki method. We will examine some of the details of how a child comes to learn to play the cello in this setting, paying particular attention to both the roles of artefacts and the interpersonal social world, and how they work to support the child's musical development. Our focus is on how the child, within this setting, develops a series of interlocked abilities to control and structure a musical performance, and how these abilities go hand in hand with a deepening sophistication and perceptual sensitivity to the many modalities and possibilities of performance and the world of music.

First, let us discuss the cello and the requirements it imposes. The cello – like all fretless stringed instruments – is a demanding teacher. A central set of skills the child must develop is a practical understanding of the cello itself. Making any kind of music on such an instrument requires not just balancing or holding the instrument in a suitable way but also holding and manipulating the bow.[16] Controlling several interrelated degrees of freedom of movement at the same time is required in order to produce a reasonable sound. Playing even a basic melody on the cello is almost certainly a harder challenge than the more usual beginners instruments of recorder, harmonica, or guitar. Finding and sounding the pitch of any particular note (intonation) is a task requiring considerable perceptual sensitivity and motor dexterity. Making this into a musical sound demands not only the skills of controlling the bow, holding the instrument and finding the right notes to play, but also the development of a more sensitive musical ear. That is, perceptual development and growth are required alongside considerable control of movement and posture.

One characteristic way that the Suzuki method seeks to manage the difficulty and complexity of producing a note with good intonation is by placing lines of colored tape (virtual frets!) on the fingerboard of the instrument (fig. 1). These are, in the first instance, placed where the major second, third and perhaps fourth degrees of a major scale would be – considering the open string as the root note – and serve as visual guides to where the student should place their fingers in order to produce good intonation. In this way, the child can make use of the visual cues (tape), so to orient their growing abilities to find the desired pitch of a note through a proprioceptive familiarity of how one´s fingers reach for that note in combination with what their musical ear seeks to hear. The tape then serves as a 'scaffold' from which more refined perceptual and active possibilities are developed. Eventually, the visual clues become superseded by the child's proprioceptive knowledge of where to put their fingers alongside the growing ability to *hear* whether a given note is in tune. At this point, the colored tapes are removed from the instrument. This takes some time!

[16] Holding the bow and actually bowing is a task that may take years to accomplish.

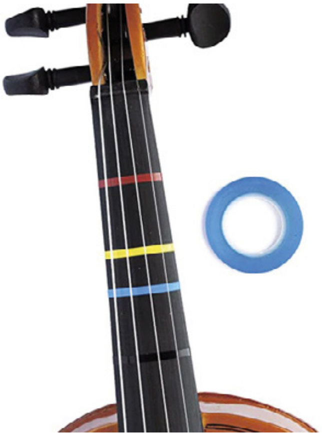


FIGURE 1

However, the ability to play with good intonation is, if anything, an easier and more natural task than learning to use the bow in such a way as to produce a consistent and well-defined musical sound on a string. This requires, in the first instance, significant attention to posture and movement in order to develop the habits and inculcate a set of skills whereby the child becomes progressively able to control the fluid and seemingly effortless production of musical sound. Learning to do this effectively requires a significant development of motor-control and postural sensitivity on the part of the child. The enveloping socio-cultural setting, at least with the Suzuki method, plays a central and explicit part in developing these abilities.

The Suzuki method depends on the cultivation of a special nurturing relationship between teacher, child, and parent (or caregiver).[17] The parent is introduced to the Suzuki method and theory, and is often encouraged to attend lessons. A guiding idea of this relationship is to bring the child, and importantly the parent, to a gradually deepening sensitivity to musical practice and to musical performance. Part of the reason is to develop a supportive and knowledgeable background to aid the growth of skilled practice and refined perception in the child. The cultivation of this background helps the parent to support the child's

[17] A note on the method cited from the cover of a current Cello School book reads that 'The Suzuki Method involves the student, the trained teacher, and the parent. Parents work with teachers to create a fun, nurturing environment for learning by attending lessons with their child, serving as "home teachers," and playing music at home.' (Suzuki, 1991, back cover)

home practice time which is crucial to the growth of their abilities to gradually refine their skills more autonomously.

Implicitly, there appears to be something of a necessary distribution of cognitive labor between child and parent.[18] While the (young) child is deeply engaged in the business of holding the bow just so (e.g., respecting the balance point), or resting the cello on a particular point on their chest and left knee, the parent is often busy taking notes to help prompt later practice sessions at home.[19] It is not unusual for the Suzuki student to begin their learning at four or five years of age, so it is certainly the case that the parent needs to take substantial responsibility for explicitly remembering what the child is supposed to be doing, while the child is concerned more with the physical requirements of the cello. As the child's sophistication grows, they are able to take over more of the explicit memory burden making (mental) notes from a weekly lesson in order to control the structure of home practice for themselves, and indeed taking more responsibility for their own targets and goals.

Although the Suzuki method emphasizes how playing music and the production of tone precedes any deep musical theory, at least as presented to the child, as the child learns to inhabit the physical constraints that allow the production of musical sound, they also have to develop their abilities to understand and to hear music. These aspects are crucial and eventually come to be accompanied by substantial theoretical knowledge. Cello lessons are therefore often accompanied with music reading and eventually theory classes, which introduce children to musical notation – alongside many central musical concepts about how to modulate and control performance. Another factor of importance is the "class conjunto": a joint class where the students play in a mini-cello

[18] This could be thought of as a case of distributed skill acquisition. See Sutton et al., 2020 for an account on distributed memory and socially distributed remembering.

[19] This can also be looked at as a sort of social division of memory. Whereas the child is (implicitly) learning how it feels to hold the cello and the bow in a way the musical occasion demands – often a process of implicit practical understanding of the posture of the body – the parent may be taking notes of key points, such as what should be remembered or picked out for detailed attention in the practice sessions at home between the weekly sessions with the cello teacher.

orchestra and where children begin to learn about the difficulties and joys of playing with others and also working towards performance in front of an audience.[20]

Playing the cello and certainly anything approaching mastery requires a great depth of skilled performance. The knowledge required is not just of the theoretical type we have just mentioned, but prior to this, and also grounding it, a practical knowledge of skillful practice involving an attunement to the requirements of the artefact mediated production of sound. The acquisition of the requisite skills is – as we have just described – heavily reliant upon the production of custom environments, a densely socially scaffolded process of acquiring specific skills and, not least, the gradual physical attunement to performance, e.g., adaptation of muscles, sense of bodily posture and even how to manage emotions through performance. These skills are very aptly described as being situated-embodied practices. They require the child to build a sensitivity to certain tools, the cello, the bow, the resin, and how they interact, as well as a deep appreciation of the possibilities and constraints of movement of their own bodies with respect to these implements. Playing well will also require significant muscle control and an appreciation of posture to the point that it becomes a sort of transparent habit. It will eventually also require a working understanding of related and sometimes less practical domains such as more theoretical knowledge (e.g., scales, arpeggios, the cycle of fifths, etc.) and other forms of practical understanding, (such as how to read music, relate to others in the orchestra, play at an appropriate volume, etc.), and how all of these too relate to the production of sound (e.g., the production of vibrato, using the bow).

A dense set of practical skills, theoretical knowledge and situated-embodied practices all have to come together to allow a good performance. In performance, it is gradually no longer the production of an individual note, or successful body posture that is uppermost in a student's mind. These foundational features need to be in part taken for granted and are handled by the habits of skilled practice. It is the performance of quality that is the ultimate goal and eventually many of these hard-

[20] In fact, at least in the music school of one of us, children are encouraged to make recital performances for groups of parents and family more or less from the beginning of their musical journey.

earned skills will become, and need to become, transparent in action, so that the student can think of the performance in more abstract terms. The student instead learns to think of musical phrases, or passages of play, or how a whole performance can be subtly modulated and approached in a number of ways, about which the performer can make conscious choices in order to evoke a different emotional response. Indeed, the development of such practices are intimately related to a growing sophistication and perceptual refinement on the part of the student. This perceptual growth can be seen as moving outwards from the production of a single note to an expansion across a number of temporal and spatial scales such that the child's mind is able to focus on more refined, more temporally extended, more abstract and more emotionally affective aspects of performance.[21]

We have introduced this vignette at length to focus on a particular area of human activity which we think is neglected when we consider mental depth. We believe that it is largely these sorts of practices in which human mental depth inheres. It is essentially a set of skilled and situated practices, dependent on a particular artefactual culture and a variety of social supports, practices and interpersonal relationships. The type of mental depth that is produced can be highly resilient but is also highly situated and depended on particular environmental supports and extended processes of education and enculturation. It is not however in any useful sense illusory or confabulated. In what follows, we will be taking this sort of skilled practices as an archetype of human mental depth.

4. The sources of depth of the skillful mind

We will now demonstrate how it is possible to theorize the acquisition and refinement of skilled mental depth in terms of some particular reference points in contemporary cognitive science.

[21] Although we do not have the space to really develop these ideas here, it is possible to cast the child's developing skills in a more general temporal framework, proceeding outwards from what developmental psychologist Margaret Donaldson call the "point mode" of the here and now to ever expanding vistas of time in which a performance is located (Donaldson, 1979, 1992)

4.1 Mental depth and the rich interpretation of the situation

Skills, while being interactive and dependent upon environmental props, also require something that the agent brings to the situation. A skilled agent may depend on proximal tools and what these tools afford, but it is undeniable that agents also bring ways of acting in and seeing to that situation. These ways of seeing and acting, these abilities to pick out the unique affordances that only a skillful agent can see are in need of explanation. Our account will make sense of this ability to see and develop sensitivities to new affordances by giving an account of the neuronal contribution to the experience and ongoing activity of mental depth, especially as it appears in the cultivation and exercise of human skillful action. We will first present an account of skill growth drawing from the predictive processing framework (Clark, 2016), before returning to our account of the embodied and situated nature of skills.

Let us begin by reviewing some Predictive Processing basics. Two core elements of predictive processing are especially relevant for our purpose. The first is the *generative model*, whose main task is the prediction of sensory signals, and the second is the *precision-estimation mechanism*. The generative model is a unified body of acquired knowledge based on previous experiences. A central thesis of the predictive processing account is that the way perception works is by predicting bottom-up sensory cues drawn from its best models of what is likely to be causing them. Perception is an active process, and the brain contributes to this activity. The prediction of sensory input drawing from the statistically salient history of the agent is a risky process, that is, it could easily go wrong. That is why there is a second feature in the predictive processing framework that works with the generative model, namely the precision-estimation mechanism. The central idea is that, besides the prediction task, the brain assigns a probability to the source of information given its estimated degree of certainty or uncertainty. The result of the interplay between these two mechanisms is both an optimization of top-down predictions, or *priors*, together with an optimization of the precision-estimations that determine the probability of bottom-up prediction errors in the processing hierarchy. In other words, mismatches between expectation and input (i.e., prediction errors) are propagated 'forward' in the system where they serve to further refine the top-down predictions. Supporting this *top-down* or *knowledge-driven* prediction, we can attend for instance, to the "hollow face illusion" Clark (2015, p. 3767). This illusion is produced because we are used to convex faces in our daily experiences, that is, we *know* faces are convex. According to the predictive processing account of perception, these experiences are accumulated, contributing to our *generative model* of statistically salient experiences. The illusion happens when we are presented with a screen with a concave mask. One way of explaining this effect is by assuming that the expectation of convexness is so strong that we cannot help but experience certain face-like input as convex. This means that we expect faces to be convex, or in more accurate terms, we have a "deep sub-personal 'expectation' of convexness" (Clark, 2015, p. 3767). In what follows, we use these elements (i.e., the generative model and precision estimation) which are shaped by our prior actions and histories of interactions through the development of skillful practice, in order to explain their contributions to mental depth.

Once we reject an intracranialist approach to cognition (more on this in the next section), the brain emerges as part of the larger ensemble that is the embodied cognitive system. This system is in constant interaction with, and dependent upon its changing environment through a series of looping interactions. According to our preferred view on predictive processing, the predictive brain is an extra element of situated embodiment and not a replacement for it (cf Hohwy, 2013). One way of describing simply what happens when we perceive something according to predictive processing, is that "we see the world by (if you will) guessing the world, using the sensory signal to refine and nuance the guessing as we go along" (Clark, 2016, 43). As we have just seen, the process of trying to guess or accurately predict worldly states is accomplished sub-personally by the rich mechanisms afforded by the predictive brain. However, we also act on the world in order to optimize precision-estimation, or in other words, to render it unsurprising. This is achieved by means of what has been called *active inference*.[22] As Fabry (2018) writes, in active inference: "embodied actions bring about changes in the available sensory input so as to confirm the accuracy and adequacy of top-down predictions. On this construal, any type of bodily movement – from ocular-motor adjustments to locomotion – has the potential to confirm the best probabilistically generated predictions" (p. 2486). Notice that this role of embodied action in perception is in tune with what we saw in our previous discussion of learning to play the cello and especially how this relates to a growth in perceptual and related conceptual abilities. We will shortly develop this idea further by relating it to an account of how perception is constituted by a practical knowledge of sensorimotor contingencies, or the structured nature of worldly affordances (Noë, 2002).

The picture we offer is the following. The embodied agent generates internal dynamics based on their prior history of interactions mainly in virtue of neural networks that compose the generative model;[23] these endow us with sub-personal mechanisms that infer the likelihood or uncertainty of sensory regularities based on what has been previously encountered. Importantly, we act on the world in order to adjust sensory information with prior predictions. Perception is thus constrained and contextualized by prior action and the history of the agent. This constraint and contextualization forms the neural basis of our experience of mental depth, by allowing us to *recognize* complex and affordances for action in the world, especially in situations in which we have developed skillful mastery.

[22] Hohwy (2013) distinguishes between perceptual inference (i.e., accurately predicting worldly states) and active inference (i.e, acting on the world in order to optimize the precision estimation task). As he notes, both of them are computationally similar, since they are coordinated by the optimization of precision estimations. However, he claims that they have a "different direction of fit" (p. 178)

[23] It has to be admitted here that despite the widespread view that the idea of generative model can be grounded in the structure of the brain, the precise neural implementation of this model, and indeed whether real brains are predictive processing systems remains controversial. Our discussion then of predictive processing in neural terms is of a still uncertain empirical hypothesis, and one that is the subject of ongoing disputation and the search for empirical evidence.

To illustrate this, let us return to our cello example. Practicing the cello requires the slow development of a series of sensitivities toward the instrument including how to hold the instrument appropriately, how to use the bow, and how to find a note. For a more skilled practitioner some of these aspects become increasingly facile and transparent and the performer can concentrate on other dimensions of performance. If we apply the concepts that the predictive processing framework offers us, we can explain what is at stake in the following way.

Generative models or "the generative model" are the proposed neural realization of what we bring to the situation in order to appropriately respond to it.[24]

Through the tacit practical knowledge made available through a much-refined generative model, the adept cello player sees possibilities for controlled action that are not available to the novice. Put another way, the adept cello player does not see, hold, or hear the same instrument

[24] We refer to generative models here but note that the term generative model is the more usual and canonical terminology. The terms generative model implies that the whole brain can be seen as one multi-level (hierarchical) and multi-scalar interrelated model of the causal structure of the world with which an organism interacts. Karl Friston´s use of the term generative model also makes a connection to the free energy principle, which sees predictive processing as fundamentally being a way of reducing free energy (See Wiese & Metzinger, 2017 for a primer on free energy principle and how this interacts with hierarchical models). Another idea – and more important to the context of our discussion – is that the generative model with the highest posterior probability, or, put another way, the "winning hypothesis" of the brain about the current causal structure of the world can be used to explain the content of consciousness at that moment (e.g., Hohwy & Seth, 2020). We take both of these important, and possibly revolutionary ideas to be interestingly controversial but largely beyond the scope of this paper. The use of the term 'generative models' indicates then that one could hold a predictive processing account of mind without subscribing to either of the two just mentioned propositions with the conceptual implication that it is possible to hold a form of predictive processing view without assuming that all "generative models" in the brain need be integrated into one model. The notion that there may be several predictive processing models in the brain may also comport well with some other theories of consciousness such as Dennett´s (1991) multiple drafts theory of consciousness and we draw the reader´s attention to a recent paper that assays just this possibility (Dołega & Dewhurst, 2021). Indeed, since predictive processing currently appears consistent with a number of different theories of consciousness this could be seen as a limitation of the view, at least as a means of explaining consciousness (Schlicht & Dołega, 2021). Thanks to an anonymous reviewer for encouraging us to clarify this further.

as the novice. Hours of practice and training are reflected in the models that the adept brings to the situation and thus, the affordances that they perceive. The process of practicing can be understood as a process of refinement of generative models. Precision-estimation mechanisms contribute to the refinement that takes place during the process of practice. The propagation of errors and constant attunement contributes to developing and fixating more accurate and refined strategies for the controlled expression of music. Importantly, these processes take place in a sub-personal fashion, giving rise to a form of 'self-supervised learning' (Clark, 2016, p. 18). The idea is simple: by constantly trying to infer the sensory signal via the generative model, our history walks with us in a subtle way, gradually shaping our experience of the world and the depth that we find there. It is the deep history of our skillful interactions – as we achieve mastery of the sensorimotor dependencies in any given domain or type of situation – that determine how we see the affordances therein. The precision-estimation mechanisms contribute to the process of learning and achieving the mastery of skills, by contributing to the accuracy of our skilled gestures in action.

One other aspect of the importance of generative models here is *Active Inference.* Active inference emphasizes how probabilistic beliefs are not only changed by optimization of the generative model, but also through acting *on the world.* One way of controlling the variations of sound produced by bowing is by subtly changing one's grip on the bow. Performing an action in one way or other changes incoming sensory data, and thus allows us to refine different aspects of a performance. A central way in which the predictive processing idea meets up with skilled practice is in how it allows us to give an account of the integral role of movement and action in allowing one's models to produce the optimal grip upon the world in order to produce skilled performance.

We would like to note again here how developing a skill such a cello performance requires a significant growth in perception; a greater sensitivity to the richness and multiple modalities of the world and how we can act upon it (Donaldson, 1979, 1992). One way to think about this is in regard to the two other ways in which predictive processing theory seeks to explain our cognitive abilities. It is said that perception realized through generative models is "richly world revealing" (Clark, 2012). A central idea of predictive processing is that a generative model is a model of causal structures in the world. We do not perceive surfaces as such but the interplay of causal forces that "explain" what our sensory systems receive. Such a view allows us to naturally locate the development of perceptual sophistication and how it is tied to the development of skillful action. It is not just that we develop more sophisticated skillful practices, but these practices reveal to us a growing sense of the intricacy and interplay of forces in the world beyond us. Moreover, predictive processing proposes a much more unified account of perception and action than classical cognitive accounts. These two factors can be very naturally applied to music where it is through the development of skillful practice that much of the depth of music and musical structure is revealed. This close intertwining of action and perception allows us to explain how musical sensitivity grows. It seems natural to link this to the growth of generative models and to use these to account for the expanding richness and temporal range of our perceptual abilities. The growth of sensitivity in the performer can be explained in part through the growth of perceptual depth and its tight reliance upon situated and embodied skillful practices. The predictive processing account puts on a firm foundation what the brain brings to the acquisition of deep skills.

With this in mind, we can go back to Chater's account of mental depth. If our account is on the right track, this means that the claim that there is no back- ground processing is not strictly true. Our claim is that there *is* indeed a form of background processing, which can be accounted for in terms of the hierarchical predictive knowledge with which we perceive and act on the world, and that contributes to eliciting the experience of mental depth. In this respect, depth is partly explained by our prior associations and experience. However, this background processing should not be conceived of in a classically cognitivist and internalist way. As will become clearer in the next section, we understand predictive processing in embodied and situated terms, where the brain is part of a wider embodied system.[25]

[25] See, for instance, Clark, 2015, 2016; Fabry, 2018; Gallagher & Allen, 2018, to mention some. Though the level of embodiment varies, all of them reject the view where predictive processing is simply a matter of the working brain.

4.2. Situated and embodied depth

Having put this predictive processing account of cognitive architecture in place, we now foreground how mental depth is necessarily situated and embodied and substantially relies upon, and indeed is constituted by our skills. Mental Depth is not something best understood as purely internal, or that happens outside of the conscious mind nor is it a narratively generated confabulation. Our claim, moreover, is that mental depth is not simply the result of neural activity. Rather it emerges through the development of specific practices scaffolded by the social world and through deep enculturation. The development of skilled interactions and artefactual manipulations allow us to make use of the potentialities of the rich material environments in which we are embedded. These and that actively contribute to the feeling of depth and coherence. In the previous section (4.1), we focused on the neural contribution to mental depth in terms of predictive processing. Here we will explain its situated and embodied character and provide the theoretical framework that allows us to ground our account of mental depth in skillful action.

Let us begin by observing that the folk image of our psychological lives is that the mind (i.e., one's thoughts, desires, memories, etc.) is located somewhere in the head. This idea receives further support from much cognitive neuroscience, according to which mental and cognitive processes are implemented *only* by the brain and the central nervous system. In this way, according to this folk position, mental processes take place somewhere *in* the brain and the central nervous system. We may call this view *cognitive intracranialism*, and we can locate Chater's account within the space of intracranialist views. Recent approaches in cognitive science challenge cognitive intracranialism by expanding the cognitive realm so as to include the agent's body and the skillful interactions in an environment (Newen, De Bruin, & Gallagher, 2018).

Our account of mental depth rejects *cognitive intracranialism* and instead conceives of cognition, skillful action and mental depth as all strongly embodied and embedded. By strongly embodied, we mean that cognitive processes in- volve the body acting in and on the environment (Clark, 1997; Gallagher, 2005). By strongly embedded, we understand that some mental and cognitive processes are the result of the integration with states and processes found in the environment (Menary, 2007). We conceive of the environment not simply as the physical environment but also a structured social, artefactual, and cultural environment. This is captured by our reference to the *situatedness* of cognition.

In order to present how our situated approach to mental depth contrasts with Chater's view, we will review Noë's (2002) reply to an illusionist challenge faced by the standard cognitivist approach to (especially) visual perception (Black- more, 2002; Dennett, 1991). The challenge is the following. We know that the information received by the visual system is fragmented and discontinuous, e.g., we are susceptible to change blindness type illusions. Yet, our visual experience of the world, in a phenomenological sense, generally appears coherent and continuous. However, it is believed that we have "a richly detailed picture- like experience of the world, one that represents the world in sharp focus, uniform detail, and high resolution from the center out to the periphery" (Noë, 2002, p. 2). To explain how that could be the case, one approach is to argue that the brain needs to 'fill in' the missing sensory details. However, there is no definite evidence that the brain actually 'fills in' all the gaps. (Dennett 1991, pp. 344–356). The consequence is that despite the *apparent* continuity of conscious- ness in experience, consciousness might in fact be discontinuous. Similar to Chater's criticism regarding mental depth, there is thus an asymmetry between common beliefs concerning perceptual experience and what perceptual experience really is.

One way of making sense of this situation is by claiming that we normally fall prey to a continuity illusion (Dennett, 1991). We are victims of an illusion concerning the continuity of experience or the character of visual consciousness. We are thus wrong when it comes to what we take our experiences to be. Resisting this option, a strongly embodied or enacted view on cognition opens up a different possibility (Noë, 2002). The main idea is that the illusionist challenge is built on an incorrect view on what the phenomenology of perceptual experience is.

A central thesis of the embodied approach to cognition is the continuity of perception and action (Clark, 1997; Hurley, 1998; Noë, 2002). This thesis argues against the standard cognitivist approach according to which perception is a passive process, and sensory stimuli are received and processed in order to give rise to a behavioral output. A strongly embodied or enactive approach to cognition, conceives perception as intimately linked to the action of an organism in its environment. One way of articulating this continuity is by considering that perceptual experience is in fact constituted by sensorimotor expectations acquired and developed through the exploration of the environment (Hurley, 1998; Noë, 2004). This rejects the standard construal which, borrowing from Hurley (1998), is committed to the "classical sandwich model", that is, the claim that cognition is like the filling of a sandwich which is contained by perception and action. Instead, on the strongly embodied construal of cognition, perception and action are intimately connected to the extent that perceiving *is* a form of acting.

These points call into question what the illusionists took perceptual phenomenology to be: we do not perceptually experience the world as a whole in a "snapshot" but we experience having access to the world in virtue of our sensorimotor abilities Noë, 2004). In other words, perceptual experience is not best understood as a rich perceptual world which is presented or given to our sensory channels, but as a skilled cognitive engagement. Take, for instance, the experience of visually perceiving a wall. We don't experience the whole of the wall's surface but we "experience the wall as present and you experience your- self as having access to the wall by looking here, or there, attending here, or there" (Noë 2004, 4). This offers a solution against illusionism: consciousness is not continuous in the standard "snapshot" sense, but this does not entail that it is discontinuous, and thus we are constantly prey to an illusion. What this shows is that we need to move beyond the standard construal of perceptual experience.[26]

Our strategy against Chater's challenge to mental depth follows a similar dialectical structure. We resist the claim that mental depth is an illusion, but our resistance only succeeds if we adopt a different and more plausible phenomenology of mental depth.

[26] This also allows the explanation of the *perceptual presence* of objects or scenes despite them not being entirely modally present. The idea is the following: when you see a bottle in front of you, you perceive the bottle as being entirely present, despite the fact that from where you stand, you cannot see the back of the bottle. For more on this see Noë, 2002, p. 9.

To make our case, let us briefly recall Chater's challenge. Chater challenges the folk/standard conception of mental depth, according to which our inner mental life is as deep as it is depicted in literary novels and in the assumptions of folk-psychology. The challenge is grounded in the fact that according to Chater's view there is no background processing – a claim we have just challenged – and our mental life is in fact sketchy, gappy, and highly incoherent. As we have seen, Chater claims that much of the detail of our mental imagery is filled in as and when needed in a manner that Chater describes as "the cycle of thought." The consequence that Chater draws from this is that our experience of mental depth is thus illusory. But is this the only option? We don't think so. In fact, once we endorse a strongly embodied and embedded approach to cognition, we can offer a new account of mental depth and its phenomenology.

As we have stated earlier in this section, our account of mental depth begins with the consideration that cognition is largely a form of situated skillful activity. In other words, cognition is largely a matter of the embodied manipulation of external structures (Rowlands, 1999). This means that it is in virtue of learning such manipulative skills and the acquisition of social practices that our cognitive abilities are transformed and extended (Menary, 2007, 2012, 2018). Cognition is thus not solely an intracranial activity, but it includes different patterns of skillful action. Understanding cognition requires us to pay attention to the different "skilled gestures" (Clowes, 2019; Andrada, 2020), that is the skilled movements and manipulations that allow us to act and interact with the mate- rial culture surrounding us as well as with others.[27]

Taking seriously such a strongly embodied and situated approach calls for a reconfiguration of our notion of mental depth. It requires us to see that mental depth is not an entirely inner matter, constituted by hidden causes and background processing, but depends upon the kind of access we have to our environment. Especially insofar as our access involves skilled manipulations and sensorimotor interactions. In other words, depth phenomenology is not simply the result of neural activity,

[27] Extended cognition so understood is thus "a theory of *access*" (Krueger, 2014, p. 5). This approach to extended cognition is less ontologically committed and thus less controversial given that cognition is extended, that is, cognition includes extra-organismic elements only when engaged in sensorimotor dynamics.

but the way the neural activity is embedded in a sensorimotor dynamic involving conceptual growth and environmental exploration.[28] This forces us to reconsider what Chater takes the default position concerning the phenomenology of mental depth to be. It is not that we experience ourselves as deep characters with a rich inner and largely hidden mental life, but we discover ourselves as deep in virtue of our situated activities in our active engagement with and responsiveness to the environment. In other words, we are not under the illusion of depth, but depth is not as folk psychology traditionally took it to be.

The phenomenology of depth begins by noting that it is through our activities, exploration, and developing sensitivities to worldly events and occurrences that we discover depth. We find depth in the course of our worldly activities through the sense of multiple ways of acting upon and thus thinking about an object, tool, or situation. We create and experience depth especially in the exercise and mastery of skillful practices, or sometimes through the observation of those practices being performed by others. We find depth in worldly inter- actions and practices. Just as our discussion of learning to play the cello suggested, mental depth comes from our ability to attune to and/or respond to these sort of rich practices and structures as we find them in the world.

One way of noticing how our skillful practices in the local environments con- tribute to the phenomenology of mental depth is by noting what happens when we are deprived of access to some aspect of the enveloping material and social environment upon which we usually depend. In such circumstances we can experience the loss as a sort of deprivation of parts of ourself, or, put another way, as a loss of mental depth (more in the next section). That is why we con- sider the kind of access we gain in virtue of our situated skillful practices as in fact at least partly constitutive of the experience of mental depth. This access, though

[28] It should be noted that this situated and embodied aspect of mental coherence connects with Merleau Ponty's notion of 'primordial depth' which is not the result of cognitive calculation but something that is gained and achieved by an agent in virtue of being situated in the world. Depth is thus experienced by a "subject involved in the world" (1962, pp. 256–7). Merleau-Ponty writes that depth "is, so to speak the most 'existential' of all dimensions, because... it quite clearly belongs to the perspective and not to things... It announces a certain indissoluble link between things and myself by which I am placed in front of them" (1967, p. 256; quoted in Bredlau, 2010, p. 413).

it can be, needs not be conscious, and for the most part, will not be the focus of our attention and conscious thought, at least to those practicing the necessary skillful activities. However, there will be a particular phenomenology associated with this constant access which is afforded by our embodied manipulations.

Moreover, conceiving of mental depth as situated and embodied can also help us understand mental coherence in a way that does not require the sort of inwardness of mental life that Chater challenges. Take for instance what Reed's (1993) writes recalling his experience with his patients suffering from traumatic closed head injuries with severe brain damage. Despite their failure to organize coherent skilled behavior (e.g., preparing coffee or pouring water), they were able to preserve certain coherence through their actions (e.g., they mistook liquids but they did not pour liquids on a plate, nor they did put solids in a glass). This can be explained by the fact that they preserved certain basic affordances learned in virtue of their previous interactions within their local environments, and their local environments partly contributed to their mental coherence.

Summing up, the depth of human mental life is based in our skillful practices in the world of the sort we have just assayed in our discussion of the acquisition of skill of playing the cello. Although the acquisition and maintenance of such a skill is in many respects particular and *sui generis*, skills have a sort of general structure which can be used to articulate much of what is specific to the human mind and our particular form of mental depth. Depth is acquired and sustained through the development of skills that involve the elaboration of specific abilities which are often or usually highly dependent upon the tools and artefacts of the human-made environment. We thus are mentally deep not because of inscrutable hidden processes, but because of the way we are able to respond to the world. This situated skillful depth is not illusory, and neither is it in any simple sense inner. Rather, in order to understand both how it arises, and what is its essential nature, we need to understand how it intrinsically depends upon the situated nature of our being in the world.

We thus conclude that depth is not an illusion that results from our conscious or unconscious narratives and confabulation, but it emerges and is present in our skillful practice. Our experience of depth on this model is deeply related to our ability to produce complex, structured, and refined activities. These abilities are rooted in the history of our development of skills and largely resides in the prior associations, expectations (generative models), and experiences which are built-into the architecture of our perceptual and cognitive embodied mechanisms. The conclusion of this section is that the deep wells of mind stem from the twin pillars of the generative models of the brain and the embodied and situated nature of our cognitive systems. The depth of the human mind is not hidden away, nor a confabulation, but emerges in the rich interactive nature of the practices of mind in the world.

5. The real depth of the interactive mind

Chater argues that "Our brains are, then, relentless and compelling improvisers, creating the mind, moment by moment" (p. 220). We have argued this is only partly right. It is true we can be masterly improvisors, but these abilities emerge from our slowly acquisition of deep and world-revealing knowledge. The skillful mind is created and maintained over durations that are deeper and more extended than momentary improvisation. The case of learning to play cello has helped motivate and illustrate our proposal. It allowed us to present the means by which human beings acquire deep knowledge and the capacity for skilled, involved action. In the previous section, we have argued there are two sources of the depth of the human mind. The first source of depth is the hierarchical predictive knowledge we bring to those situations, as PP teaches us, our brains can be considered as a collection of highly integrated generative models or, more canonically, just one deeply-integrated generative model. The second source of depth is our rich and skillful embedding in the world. We are (often) skilled and profound manipulators of the environment we find ourselves in. The development of skillful performance can be understood as a gradual elaboration of skills, and with it a gradual expansion and refinement of our perceptual capabilities. By linking this account to predictive processing, we have also argued that these skills are grounded in the real causal structure of the world, and our abilities to intervene in it. As in our example of the child's growing sensitivity to music and through their growing abilities to control musical performance, skills are slowly appropriated and mastery gradually developed. Learning to play the cello is a process of ever-deepening mastery but such skilled mastery applies across the range of our worldly interactions. Skills developed in one context can (often) be deployed in an increasing range of related scenarios, i.e., learning to play vibrato in a simple piece can later be applied to many other pieces of music. Thus, the novice is able to achieve more sophisticated and resilient skills of interpretation and performance.

However, one of the virtues of our proposal is that it also allows us to explain how and why mental depth can also be lost, or at least impeded, when our skillful practices become untied from their moorings, in the sorts of situations and contexts in which they were first developed. Situated skills are always relatively dependent upon context and as such are neither infinitely adaptive nor instantly redeployable, at least in short time horizons. Take, for instance, the reported experiences of teaching during the current COVID 19 pandemic via platforms such as Zoom or Google Meet. The experience of many educators living through the pandemic and attempting to rapidly convert teaching materials and practices to online platforms was experienced by many as a losing of capacities – being unable to easily interact with students, having that interaction de-naturalized. We suggest that the embodied repertoire of skills that had been built after years of teaching in a classroom, were suddenly interrupted as many adept teachers and lecturers discovered they could not be easily put to use in the virtual classroom. Importantly, many such skills are typically in the background, relied upon implicitly. Habits such as walking across the class to capture the students' attention became inaccessible inhibiting the abilities of many experienced educators to engage with students in the same way as before. The new ecological setting of online learning thus can strip away skills that were previously taken for granted. When reflecting on our own experience and the many anecdotal reports, we believe that the dramatic change in ecological setting and resultant inhibition of relied upon skills should be described as an experience of loss of mental depth.[29] Such an experience

[29] This might have been a collective experience. Lecturers in England have been reported as "feeling out of their depth." Source: https://www.theguardian.com/education/2020/dec/03/i- feel-out-of-my-depth-university-lecturers-in-england-on-the-impact-of-the-pandemic (last accessed: January 29, 2021).

is characterized by the feeling that abilities seem much more brittle or shallow and a characteristic fatigue in performing "the same" tasks that had previously seemed unproblematic via a new medium.

Prior to the confinement many educators lacked familiarization with practices for online teaching and learning, and that can also partly explain the feeling of losing depth and the diminished capacity for improvisation and fluent action. Of course, this is not to say it was impossible to teach well using tools like Zoom etc. In fact, as time passed by, we expect many educators would find themselves able to develop new refined practices responsive to online group dynamics and anchored in the new tools. That is why, after a sudden change in the usual situation, a period of training and adjustment was needed to regain the feeling of depth. The loss of the feeling of depth reveals, we think, something real. As presential teaching is replaced by online formats, many of our skillful practices become inhibited and new ones have to be developed. The reality of depth is revealed when it is lost and sometimes regained. We think this experience can be explained in virtue of the fact that much of the background processes upon which teaching relies: the situated and embodied context in which habits and skills are embedded were stripped away. Our skills always relate to, or at least are formed, in particular environments and our cognitive systems are intimately tuned to the material culture in which we are embedded.

Summing up, our claim that mental depth is embodied and situated has allowed us to argue that depth is not illusory, but we should reconceive what both psychology and folk-psychology took mental depth to be.

Before concluding, we want to consider a possible objection to our account. One could argue, drawing from Chater's remarks, that despite our analysis of skillful practices the notion of situated depth still mental depth is apparent and confabulated. To allay this worry, let us finish by explaining why it's not the case.

There is really nothing obviously confabulated about skilled practices. We need to theoretically separate our skillful embodied interactions with tools and others, from stories we tell about them. It is important to see here that we are not denying that human minds do confabulate explanations, including the narrative explanation of themselves. But this "narrativity" is not the real source of our cognitive depth. Depth is much better located in skillful practices, including our "folk psychological" practices of interacting with others (Gallagher, 2001; McGeer, 2001; Zawidzki, 2008). Much social cognition may also be framed as a form of skilled practice and our cognitive depth tends to emerge in those self-same skilled practices of interacting with others, often against the background of a rich environment. Such 'practices of mind' again are not illusory albeit our abilities to fully give verbal account of those practices (narrative) may be, under many conditions, rather weak. As the previous cases have suggested, hu- man mental depth, both in its practiced reality and occasional absences is best understood not as a confabulation but as a finely constituted, deeply situated, and also sometimes precarious set of practices.

This supports our claim that the depth of the human mind is real but lies not in unconscious streams of thought that has been bequeathed to our folk-psychology by Freud but in our worldly practices. Chater (2018) is probably correct that there is no such thing as an unconscious stream of thought. Albeit, even here – as we have remarked – it is not quite true to say there is no background processing. The constant readjustment of the generative models of the brain in virtue of precision-estimation mechanisms, under predictive processing is indeed a form of background processing. It is the processing by which our sensitivities to the deep causal structure of the world are revealed to us through the ongoing history of our actions. This readjustment process, and indeed the deployment of active inference is a largely unconscious and background, or put another way, subpersonal, process. There is also a sense that the way it allows us to "see as", that is, respond to and represent the deep causal structure of the world, can be seen as a deep and hidden reservoir of much of our activities and mental lives.

6. Conclusion

We began this paper by introducing a recent challenge put forward by Chater (2018) to the idea of mental depth, an idea that deeply permeates our folk psychological image and plays an important role in contemporary psychology. Contrary to traditional psychology and our folk image, Chater (2018) argues that the mind is not deep and, consequently, we are operating under an illusion about the nature of our own minds and cognitive processes. In this paper, we have put forward an account of mental depth that offers a partial reconciliation with the idea that we are deep mental creatures, but that at the same time is in tune with recent scientific approaches regarding our mental architecture. More precisely, we have partially agreed with Chater in that depth of the mind is not to be found (solely) on the deep background neuronal processing of the brain. Departing from Chater, we have claimed that this does not mean that our minds are not deep. To motivate our account, we have introduced a phenomenologically informed vignette that concerns the acquisition of mental depth through learning to play the cello. And we have relied on two different sources to explain this phenomenology. First, the predictive processing framework has allowed us to make a proposition about the brain's contribution to mental depth; specifically, in the instantiation of context-sensitive hierarchical knowledge produced by predictive processing systems. Second, we have drawn from a situated approach to cognition, to argue that the source of mental depth is located in our embodied skills and the situations in which we are embedded. Our moral is that mental depth is real, but we should move beyond its traditional understanding. Our mental depth is not the result of a deep unconscious or at least not only, but it results from our skilled engagement with a structured and sustained environment that we inhabit and act upon. Consequently, in order to understand human mental depth, we need not look only in the deep recesses of the brain but in how brains, bodies, and the world come together in skilled action. We find depth through worldly engagement and that is, for us, real enough.

References

Andrada, G. (2020). Transparency and the phenomenology of extended cognition. *LÍMITE Interdisciplinary Journal of Philosophy & Psychology*, 15(20), 1–17.

Barrett, L. F., & Bar, M. (2009). See it with feeling: Affective predictions during object perception. *Philosophical Transactions of the Royal Society, B*, 364(1521), 1325– 1334.

Blackmore, S. (2002). There is No Stream of Consciousness. *Journal of Consciousness Studies, 9*(5-6), 17-28.

Bredlau, S. M. (2010) A Respectful World: Merleau-Ponty and the Experience of Depth, *Human Studies,* 33, 411–423.

Chater, N. (2018). *The Mind Is Flat: The Remarkable Shallowness of the Improvising Brain*: Yale University Press.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again.* Cambridge, MA: The MIT Press.

Clark, A. (2008) *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*, Oxford: Oxford University Press.

Clark, A. (2015). What 'Extended Me' knows. *Synthese,* 192(11), 3757–3775.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford: Oxford University Press.

Clowes, R.W. (2014). Thinking in the Cloud: The Cognitive Incorporation of Cloud-Based Technology. *Philosophy and Technology*, 28(2): 261–296.

Clowes, R.W. (2019). Immaterial Engagement: Human Agency and the Cognitive Ecology of the Internet. *Phenomenology and the Cognitive Sciences*, 18(1), 259– 279.

Clowes, R. W., & Gärtner, K. (2020). The Pre-Reflective Situational Self. *Topoi*, 623– 637. https://doi.org/10.1007/s11245-018-9598-5.

Dennett, D. C. (1991). *Consciousness explained.* Harmondsworth: Penguin Books.

Dennett, D. C. (1992). The Self as a centre of Narrative Gravity. In F. Kessel, P. Cole & D. Johnson (eds.), *Self and Consciousness: Multiple Perspectives.* Hillsdale, NJ: Erlbaum.

Dennett, D. C. (2001). Surprise, surprise. *Behavioral and Brain Sciences, 24*(5), 982.

Dołęga, K. & Dewhurst, J. E. (2021). Fame in the predictive brain: a deflationary approach to explaining consciousness in the prediction error minimization framework. *Synthese,* 198(8), 7781–7806.

Donaldson, M. (1979). *Children's minds.* W. W. Norton & Company.

Donaldson, M. (1992). *Human minds: An exploration.* Allen Lane/Viking Penguin.

Fabry, R. (2018). Betwixt and between: the enculturated predictive processing approach to cognition. *Synthese,* 195(6), 2483–2518.

Fodor, J. (1975). *The Language of Thought.* Hassocks, UK: Harvester Press.

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies,* 23(11–12), 11–39.

Gallagher, S. (2001). The Practice of Mind. *Journal of Consciousness Studies, 8*(5–7), 83–108.

Gallagher, S. (2005). *How the Body Shapes the Mind.* Oxford: Oxford University Press.

Gallagher, S. & Allen, M. (2018). Active inference, enactivism and the hermeneutics of social cognition. *Synthese* 195(6), 2627–2648.

Gazzaniga, M. S. & LeDoux, J. (1978). *The Integrated Mind*. New York: Plenum.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Hohwy, J. & Seth, A. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences, 1*(II).

Hurley, S. (1998). *Consciousness in Action*. Cambridge MA: Harvard University Press.

Hutchins, E. (1995) *Cognition in the Wild*. MIT Press.

Hutchins, E. (2011). Enculturating the Supersized Mind. *Philosophical Studies*, 152(3), 437–446.

Krueger, J. (2013). Affordances and the musically extended mind. *Frontiers in Psychology,* 4, 1–12.

Mantel, H. (2011). *Wolf hall*. Fourth Estate (UK).

McGeer, V. (2001). Psycho-practice, psycho-theory and the contrastive case of autism. How practices of mind become second-nature. *Journal of Consciousness Studies, 5–7*, 109–132.

Menary, R. (2007). *Cognitive integration: mind and cognition unbounded*. Palgrave Macmillan.

Menary, R. (2010). Dimensions of Mind, *Phenomenology and the Cognitive Sciences*, 9(4), 561–578.

Menary, R. (2018). Cognitive Integration How Culture Transforms Us and Extends Our Cognitive Capabilities. In A. Newen, L. De Bruin, & S. Gallagher (eds.), *The Oxford handbook of 4E cognition* (pp. 187–215), Oxford: Oxford University Press.

Newen, A., De Bruin, L., & Gallagher, S. (eds.). (2018). *The Oxford handbook of 4E cognition*: Oxford University Press.

Noë, A. (2002). Is the visual world a grand illusion? *Journal of Consciousness Studies,* 9(5–6), 1–12.

Noë, A. (2004). *Action in perception*. Cambridge, MA: Bradford Books, MIT Press.

O'Regan, J. K. (2002). The Grand Illusion and Sensorimotor Approaches to Consciousness. *Journal of Consciousness Studies, 9*(5–6).

O'Regan, J. K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24, 939–973.

Power, Michael J. & Brewin, C. R. (1991). From Freud to cognitive science: A contemporary account of the unconscious. *British Journal of Clinical Psychology,* 30, 289–310.

Reed, E. (1993). The intention to use a specific affordance: a conceptual framework for psychology. In R. H. Wozniak & K. W. Fischer (eds.), *Development in Context: Activity and Thinking in Specific Environments*. Taylor and Francis Group.

Rowlands, M. (1999). *The Body in Mind: Understanding Cognitive Processes*. Cambridge: Cambridge University Press.

Rowlands, M. (2009). Extended Cognition and the Mark of the Cognitive. *Philosophical Psychology*, 22(1), 1–19.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences, 1*(7), 261–267.

Schlicht, T. & Dolega, K. (2021). You can't always get what you want: Predictive processing and consciousness. *Philosophy and the Mind Sciences, 2*.

Simons, D. J. & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences, 9*(1), 16–20.

Sutton, J., Harris, C. B., Keil, P. G., & Barnier, A. J. (2010). The psychology of memory, extended cognition, and socially distributed remembering. *Phenomenology and the Cognitive Sciences,* 9 (4), 521–560.

Suzuki, S. (1991). *Suzuki cello school, cello part 3* (Vol. 3). Alfred Music Publishing.

Wester, D. (1996). The Scientific Status of Unconscious Processes: Is Freud Really Dead? *Journal of Ametican Psychoanalytic Association,* 47(4), 1061–1106. https://doi.org/10.1177/-000306519904700404.

Zawidzki, T. W. (2008). The function of folk psychology: mind reading or mind shaping? *Philosophical Explorations, 11*(3), 193–210.

Coding Empathy in Dialogue

Fabrizio Macagno, Chrysi Rapanta,
Elisabeth Mayweg-Paus[a] and Mercè Garcia-Milà[b]

Macagno, F., Rapanta, C., Mayweg-Paus, E., & Garcia-Milà, M. (2022).
Coding empathy in dialogue.
***Journal of Pragmatics*, *192*, 116–132.**
https://doi.org/10.1016/j.pragma.2022.02.011

[a] Humboldt-Universität zu Berlin, Institute of Educational Studies, Einstein Center Digital Future

[b] Department of Developmental and Educational Psychology, Universitat de Barcelona

1. Introduction

The possibility of dialogue is rooted in the basic and fundamental capacity of understanding the interlocutor's utterances, or more precisely what the other means. As the literature in pragmatics clearly underscores (Clark, 1996; Grice, 1957; Leech, 1983; Levinson, 1983; Sperber and Wilson, 1995), this understanding does not correspond to the decoding of the sentence conveyed. Using an example from Gibbs (Gibbs, 1987, p. 591), the interpretation of the following exchange would be impossible if we consider only the so-called "literal meaning," or if we conceived of meaning only as a property of expressions in abstraction from particular situations, speakers, or hearers (Leech, 1983, p. 6):

> Bob: Would you like a piece of cake?
> Peter: I'm on a diet.

This dialogue presupposes not only the analysis of the context in which it occurs, but also the mutual availability of specific knowledge, which allows Bob to understand from Peter's sharing of personal information concerning the issue of diet that he refuses his offer.

This dimension of meaning and understanding defines the way we communicate. In a sense, the possibility of communication rests on what we *do not* communicate, namely what is taken for granted in our discourse. This awareness has led to a concept that is becoming crucial in a world characterized by the meeting of different cultures, "cultural literacy." In commenting on the capacity of understanding a literary text, Hirsch observed that the ability to decode writing and knowledge of the definitions of words used are not self-sufficient. To be able to understand a text (or a piece of discourse), a reader (or a speaker) needs to have access to the information that it presupposes (Hirsch, 1983, p. 165). In philosophy of language and pragmatics, one of the fundamental assumptions underlying the mechanism of speaker-hearer comprehension is the so-called mutual knowledge hypothesis (Gibbs, 1987), according to which the interpretation of utterances in conversation is grounded on a set of knowledge and beliefs that listeners share with speakers (Bach and Harnish, 1979; Leech, 1983; Levinson, 1983; Schiffer, 1972).

One dimension of mutual understanding that has practically been neglected (Macagno, 2018a; Verdonik, 2010) is the lack or conflict of common grounds, which occurs when the knowledge that the speakers assume to be shared in fact is not known or is controversial. The "uncommon ground" becomes extremely important when we move from the linguistic analysis of the *products* of interactions – the utterances – to the more complex dimension of the dialogic *process* – taking into account how interlocutors detect, negotiate, and discuss the knowledge that is not shared between them. This aspect is fundamental to several disciplines, as it relates the problem of understanding with crucial issues such as intercultural communication, value comprehension, and cultural inclusion. However, how to capture when a dialogue is aimed at developing a deep understanding, and thus bringing to light and developing a ground that is common to the interlocutors?

The goal of this paper is to address the relationship between understanding and common ground, focusing on the process that allows the development of a *dialogic empathy*. From the perspective of pragmatics and communication, empathy does not depend on the amount of knowledge that a speaker holds on the interlocutors and their interests, values, and perspectives, but rather on the disposition to acquire it when it is needed. This paper proposes an operationalization of dialogic empathy, first developing a coding scheme that allows the detection of the moves in a conversation that manifest it and then showing through examples from a multicultural corpus how dialogic empathy contributes to developing a common ground between interlocutors.

2. Empathy as a dialogic dimension

Empathy, broadly defined as the ability to experience another's emotions and perceptions (Lipman, 2003), is one of the major attitudes underpinning an individual's participation in dialogue across diversity (see, for example, Barrett, 2013; Maine et al., 2019). In contemporary Europe, the fluidity of cultural identities caused by the recent migration flows requires a reconceptualization of cultural literacy as a dialogic social practice (Maine et al., 2019; Rapanta et al., 2021). Intercultural dialogue is emphasized in recent EU policy documentation as an "open exchange of views (…) (that) requires the freedom and ability to express oneself, as well as the willingness and capacity to listen to the views of others" (CoE, 2018, p. 74). Empathy in this sense is the basis for intercultural dialogue, but no clear definition of how the same can be operationalized or promoted is present.

The starting point for analyzing the possibility of communication and deep understanding, especially in a context of different backgrounds, is the concept of dialogue. However, dialogue is not the mere verbal interaction between agents: as Nystrand puts it, the dialogic nature of discourse is not the result of speakers taking turns, but rather of the "tension, even conflict, between the conversants, between self and other" (Nystrand et al., 1997, p. 8). Such differences that underlie the possibility of a dialogue can be manifested in different ways (they can be "submerged" and hidden within talk) and result in different effects on the interlocutors (Hammond et al., 2003, p. 136). Thus, the mere recognition of (potential) tension between perspectives is an important requirement for a genuine dialogue. In classroom discourse, for example, teacher-student interactions are often monological: teachers do not take into consideration the ideas that students may have, nor do they elicit or address them (Scott et al., 2006). Similarly, student-student interactions are often merely cumulative or even eristic, focusing on the development of individual perspectives, instead of co-developing and integrating different viewpoints through critique and construction (Ford, 2008; Mercer, 2004).

The "dialectic" nature of dialogue was analyzed in depth by Martin Buber, who defined dialogue as an I-Thou relation (Buber 1923/1958). A dialogic relation was regarded as "characterized in more or less degree by the element of inclusion," which presupposes that the interlocutors live "through the common event from the standpoint of the other" (Buber, 2002[1947], pp. 114–115). According to Buber, the possibility of looking at the same state of affairs through the viewpoint of the interlocutor is the essence of dialogue, which he distinguishes from other forms of "disguised monologue," in which the interlocutors simply tolerate each other, avoiding open conflicts (Shady and Larson, 2010, p. 87). This view of dialogue is crucial for understanding the balance between common ground and diversity: dialogue is characterized by the difference of perspectives and is possible because a common ground exists or is developed between interlocutors who try to understand the other side (Shady and Larson, 2010, p. 87) without necessarily giving up one's perspective (Shady and Larson, 2010, p. 83). Dialogue is thus awareness and understanding of each other's "worldviews," with all the presuppositions on which they stand and the commitments that they imply (Buber, 1999[1957], p. 103).

As stressed in the accounts above, dialogue (or rather *genuine* dialogue) rests on this process of including the other in one's own perspective, which is described in modern psychological theories with the concept of (dialogic) empathy. Empathy has been defined in different ways, mainly based on developments of Lipps' original definition as the *inner imitation of another's feelings* (Lipps, 1903), i.e., the direct activation of an emotion through the perception of another's emotion. This proposal led to two crucial different paths (Preston and de Waal, 2002, p. 2): (a) the reduction of the empathic emotion to a perceptual reaction, leading to equating empathy to experiencing the feelings of another (Elliott et al., 2011), and (b) the distinction between the detection of another's condition or emotion and one's own emotional response. The first approach has been rejected by almost all contemporary theories (Preston and de Waal, 2002, p. 4; Scheler, 2017[1954], pp. 14–16; Zahavi, 2008) as failing to trace the distinction between the cause and the possible effect, and most importantly between the self and the other, which is considered as the essential dimension of empathy as an "other-centered" emotional state (Rogers, 1980, p. 140; Zahavi, 2014, p. 102). The second approach focuses on the imitation dimension, in which the individual who is the source of the empathic emotion (and his or her emotions) is distinguished from the empathizing subject (and his or her empathic emotion), and this gap is bridged either by experiencing or understanding the other's emotion (Scheler, 2017[1954]; Zahavi, 2008), or by the cognitive understanding of the causes of another's emotion (Goldie, 2000).

The common denominator that underlies the different theories on empathy (including the ones that regard empathy as based on the experience or perception of another's emotion, see Ben-Ze'ev, 2000, p. 110), is a form of understanding of the other (Scheler, 2017[1954], p. 12). Empathy is regarded as perspective-taking (Elliott et al., 2011, p. 43), the perception or cognitive understanding of another's frame of reference, without losing the distinction between the self and the other (Rogers, 1980, p. 140). Goldie names this defining aspect of empathy as "other-orientedness" (Goldie, 2000, p. 195). In this view, the recognition of the other is combined with the narration of the other's experience (Battaly, 2011; Goldman, 2006; Zahavi, 2014): empathy (also called "empathic understanding," see Ickes, 1993, p. 591) is an attempt to understand the "inner world of another person" (Schmid, 2001).

3. Empathy as a precondition of argumentative dialogue

The importance of empathy for dialogue was one of the fundamental aspects of ancient rhetoric, where empathic understanding was conceived as the root of genuine dialogue. The core of ancient rhetoric lies in the enthymeme, which was described by Aristotle as a syllogism (an argument) with fewer premises than the ordinary ones (Gough and Tindale, 1985; Sorensen, 1988). However, as Aristotle points out (*Rhetoric*, 1357a19-22), most enthymemes are based on premises that hold only generally, as they are only commonly accepted (Walker, 1994, p. 47; Walton, 2001, p. 106). For this reason, they result in inferences that are only likely, i.e., presumptively acceptable. Enthymemes rest on a "major" premise or generalization that consists in the specification or application of a "commonplace" or warrant (guaranteeing the passage from the premise to the conclusion, Hitchcock, 1998) to the subject matter under discussion. This epistemic aspect of enthymematic premises is considered the reason for their implicit nature (Braet, 1999, p. 107): if something is presumed to be shared by the audience, there is no need to mention it. Thus, the speaker needs to weigh between the risk of taking for granted a premise that is not shared (resulting in a general disagreement) and the danger of making explicit what is not necessary (implicitly admitting that s/he does not know his or her audience).

Aristotle developed the relationship between the audience and the enthymeme when he analyzed the maxims, namely the generalizations that can be used as implicit guarantees for rhetorical conclusions. General statements (such as "Nothing is more annoying than having neighbors") can be used in certain contexts and with certain interlocutors, but not others. Thus, to be successful, the orators need to "guess" the views and the knowledge of their audience (Aristotle, *Rhetoric*, 1395b5-12), taking for granted only what they assume to be shared and acceptable (Macagno, 2018b; Tindale, 1999, p. 112). Rhetorical arguments are thus seen as essentially related to their appropriateness to the situational context, grasped by the ancient notion of "*kairos*" (at least in one of its meanings in ancient rhetoric, see Kinneavy and Eskin, 2000, p. 433; Sullivan, 1992, p. 318). In a "relativistic" rhetorical epistemology, the speakers need to take into account and ground their arguments on what is likely to be true or acceptable for a specific audience (Untersteiner, 1954; Viano, 1955, p. 281).

The rhetorical need to adapt discourse to different audiences has been developed in contemporary argumentation theories under different concepts, all of which are essentially related to empathy. One of the cornerstones of argumentation (the development of ancient dialectics) is the notion of commitment, namely the propositions that the interlocutors are expected to defend and be consistent with (Hamblin, 1970, chap. 8). Speakers are not only committed to what they say, but also to a set of propositions that constitute the background, or the presuppositions, of their discourse. The explicit (or light-side) commitments are thus distinguished from the dark-side ones, namely the unarticulated propositions that are the tacit grounds of explicit arguments or value judgments (Walton, 1987, p. 144). Dark-side commitments are crucial for understanding the deeper premises underlying the interlocutors' viewpoints (Johnson, 1975). Unless such premises are addressed, the argumentative dialogue cannot address or undermine the other's view, and lead to a change of perspective (Gilbert, 1997). For this reason, empathy – intended as perspective taking (Gehlbach, 2004) – was viewed as the precondition of argumentation, as it represents "deep understanding," the ability to put oneself inside the interlocutor's position in an argument (Walton, 1992a, p. 255) and discover the values and the assumptions that are fundamental for understanding why a certain viewpoint was endorsed (Gilbert, 1997).

4. Operationalizing empathy in coding educational dialogues

As mentioned above, dialogicity is a central notion in education. The improvement of dialogic interactions in classroom and among students has been the focus of many studies in the last decade, aiming at promoting dialogic learning (see for example, Dawson and Venville, 2010; Erduran et al., 2004; Webb et al., 2014). From an educational perspective, dialogicity is a specific communicative attitude, corresponding to being open to different points of view (Scott et al., 2006, p. 610) and engaging with them (Howe et al., 2019). Grounded on a Bakhtinian framework, dialogicity is regarded as an attitude promoting and necessary for deeper understanding, as "to understand another person's utterance means to orient oneself with respect to it, to find the proper place for it in the corresponding context" (Scott et al., 2006; Voloshinov, 1986, p. 102). The awareness of other "voices" or perspectives is an essential educational dimension, as it augments students' perspectives, leading to an "expanded repertoire" (Wegerif et al., 2019, p. 81). The crucial educational concept of dialogicity corresponds to the dialogical manifestation of empathy that we described above. Dialogicity is "to partially inhabit the positions of others," understanding not only what is said, but more importantly the reasons and the cultural context underlying it, and the possible attitude of the speaker (Wegerif et al., 2019, p. 82).

The problem of coding dialogicity thus corresponds to the challenge of coding dialogic empathy. This endeavor has been addressed by some studies in the field of education and conversational analysis, leading to distinct types of approaches. The first focuses on the analysis types of moves that are inherently oriented to the other viewpoint, and is represented by the notion of transactivity (Berkowitz and Gibbs, 1983). Transactivity refers to the inclusion or confrontation of the other's reasoning in one's reply; however, it is limited to argumentative exchanges, and involves an analytical overlap between the types of moves and their actual realizations in dialogue ("tokens"). Felton and Kuhn translated the notion of transactivity into a coding scheme capturing the coherence of a speaker's argumentative move with the interlocutor's (Felton and Kuhn, 2001). In this way, they represented the different argumentatively relevant relations between turns. A distinct approach consists in the analysis of the possible communicative functions of the moves (not only the argumentative ones) performed in an interaction (Bereiter and Scardamalia, 2016). The Cambridge Dialogue Analysis Scheme (CDAS) (Vrikki et al., 2018) represents the most complete attempt of this research trend; however, the categories developed merge different analytical levels: a) the types of dialogical moves with their realizations (which can be indeed not dialogical), and b) the purpose of a move (its function) with its content.

The limitations of the existing methods for coding dialogic empathy underscore the need of an instrument that can at the same time 1) detects this dialogical attitude in different dialogues and contexts (not only the argumentative ones) and 2) distinguishes between different analytical levels through 3) a limited number of coding categories, which could allow the reliability of the coding system. The starting point is the twofold nature of empathy underscored by the literature, i.e., the attitude of *involving* the other (being *other-oriented*) and the activity of *talking* to the other. These dimensions are theoretically distinct, as the former can be dialogically manifested through the latter, but the latter does not necessarily involve the former. For this reason, we need to distinguish two levels of analysis: the structural (or textual) dimension, bringing to light how a turn or move can *potentially* affect the dialogic relation (Rapanta, 2019), and the connectedness dimension (or relevance, see (Macagno, 2019), which captures how a move *actually* relates to the other's discourse. A move can be dialogic from a structural perspective, namely it can be considered as potentially other-oriented independent of its dialogic context (Sarangi, 2007). However, it can only be considered as *actually* other-oriented if it is relevant, thus when its potentiality becomes manifested in a dialogue.

The codes presented in the sections below are illustrated through a sample taken from a large corpus of dialogue data collected in four European countries (England, Germany, Portugal, and Spain) within a European project aimed at developing a Cultural Literacy Learning Program (CLLP) focused on the development of dialogue and argumentation skills in pre-primary, primary and secondary students, to improve communication and understanding of diversity. The data used in this paper were collected in classrooms of urban and suburban schools from September to December 2019. Student-student and student-teacher interactions stimulated by text and film materials on social and cultural topics were videotaped and transcribed after obtaining the informed consent of students' parents. The transcriptions were fully anonymized, and the videos deleted after transcription to guarantee students' anonymity and protection[1].

4.1. Types of coding categories

As shown in Figure 1, the coding scheme consists of two dimensions: the types of moves (their dialogical function), and their realization in an interaction (relevance). The type of move represents the dimension of the "utterance type:" some types of utterances necessarily involve the interlocutors (their viewpoints, backgrounds, or response), while others do not. To this purpose, 7 types of epistemic moves were distinguished according to their dialogical functions: Stating, Accepting/Discarding, Expanding, Inviting, Metadialogical, Reasoning, and Metadialogical Reasoning. A further move was considered, which is not aimed at developing knowledge but rather managing the interaction, i.e., Managing. These moves were distinguished in two categories: low-dialogical (Stating, Accepting/Discarding, and Managerial), and high-dialogical (Expanding, Metadialogical, Reasoning, and Metadialogical Reasoning).

This distinction is drawn based on the "conventional effects" of the types of moves (Austin, 1962, p. 26). A move is defined as more or less dialogic (or potentially other-oriented) depending on: (a) the degree to which it opens up the "discourse space for exploration and varied opinions" (Boyd and Markarian, 2011); and (b) the degree to which it

[1] Full description of the anonymization procedure can be found in (Rapanta et al., 2020).

results in productive uptake or successful repair (Chin, 2006). Thus, a move is less dialogic when it does not result (in terms of its conversational effects) in a continuation of the dialogue by the interlocutor and does not build on the previous discourse or explore a viewpoint (as in case of Stating or Accepting/Discarding). In contrast, it is more dialogic when the other's viewpoint is a precondition of their utterance, i.e., when it is "transactive" (Berkowitz and Gibbs, 1985, 1983; Clarke et al., 2015; Felton and Kuhn, 2001; Vogel et al., 2016; Wegerif, 2019).

The distinction and ordering of the moves in the high and low dialogicity categories provide a criterion for detecting dialogic empathy, namely the use of specific *types* of moves. However, a move that is potentially high dialogic (such as inviting) can be unrelated to the topic or the previous moves, which results in a lack of *other-orientedness*. Thus, dialogic empathy is analyzed as detected when high dialogical moves are performed relevantly in a dialogue (turning the potential dialogic nature of the moves into actually other-oriented, or empathetic).

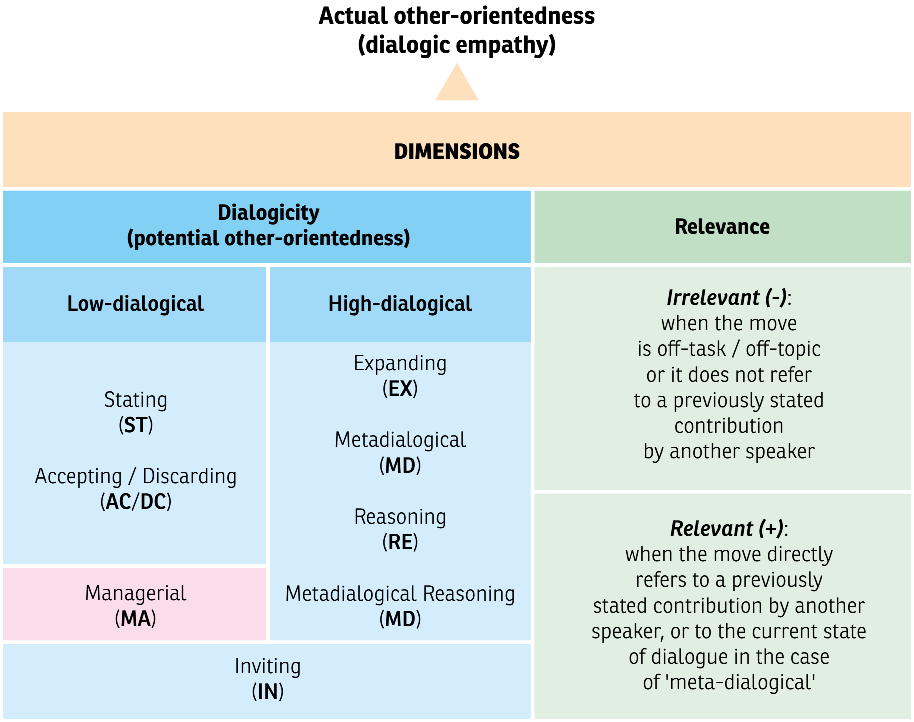

Actual other-orientedness (dialogic empathy)
DIMENSIONS
Dialogicity (potential other-orientedness)
Relevance
Low-dialogical
High-dialogical
Stating (ST)
Accepting / Discarding (AC/DC)
Managerial (MA)
Expanding (EX)
Metadialogical (MD)
Reasoning (RE)
Metadialogical Reasoning (MD)
Inviting (IN)
Irrelevant (-): when the move is off-task / off-topic or it does not refer to a previously stated contribution by another speaker
Relevant (+): when the move directly refers to a previously stated contribution by another speaker, or to the current state of dialogue in the case of 'meta-dialogical'

FIGURE 1
The dialogic empathy (other-orientedness) coding scheme

The dimensions and categories of this coding scheme are summarized below and described in detail in the annexed codebook.[2]

4.1.1. Low-dialogic categories

4.1.1.1. Managerial (MA)

A fundamental distinction in classroom discourse analysis is between "epistemic talk" (Christodoulou and Osborne, 2014), namely a dialogue aimed at the achievement of learning outcomes, and other types of talk characterized as "procedural" and "task talk" (Sarangi, 1998). Our Managerial (MA) category includes both the procedural and task-talk type and refers to the moves that are used to establish the task or norms thereof. MA moves include both the moves coordinating activities and the ones coordinating turn-taking (Table 1).

TABLE 1
Managerial moves

Example	**Explanation**
(S1) *What have you written so far?*	Activity coordination
(S2) *Let's do it like this: each one says something and then we decide which suggestion to go with.*	Turn-taking coordination

4.1.1.2. Stating (ST)

This coding category refers to "representations," namely the conveyance of information, viewpoints, or value judgments on a state of affairs or another viewpoint (Labov & Fanshel, 1977). This code includes any act of stating or asserting that something is true or false *without defending such assertion*. Stating (ST) is defined based on its dialogic effects, not on its grammatical form. Therefore, this move can also be performed through sentences that are not assertive (Searle, 1975). Table 2 provides some examples of Stating.

[2] Supplementary materials available at https://doi.org/10.1016/j.pragma.2022.02.011.

TABLE 2
Stating moves

Example	Explanation
(S1) *The text speaks about diversity.*	Advancing a viewpoint through an affirmative sentence
(S2) *But actually that is not that important, is it?*	Advancing a viewpoint through an interrogative sentence

4.1.1.3. Accepting/Discarding (AC/DC)

Any act of accepting, acknowledging (AC), challenging, or rejecting (DC) an opinion or a state of affairs put forward by another speaker, without providing further reasons and without addressing potentially problematic background values, presuppositions, or linguistic terminology, is considered an Accepting/Discarding (AC/DC) code. It can range from a simple expression of a positive or negative reaction to a more elaborate sign of agreement with another person's perspective or opinion, either through restating it or reformulating it, but without justifying such agreement (Table 3). This code includes any addition of information without the intent of making the others understand or improve their understanding of a previous move and without advancing a new idea.

TABLE 3
Accepting/Discarding moves

Example	Explanation
(S1) *You must ALWAYS follow the rules.* (ST) (S2) *I disagree* (DC)	S2 disagrees with S1 without advancing an original viewpoint
(T) *I also think that Aladdin is the main character.* (AC)	The teacher expresses her agreement; she does not put forward any new viewpoint or clarification

4.1.2. High-dialogic categories

4.1.2.1. Expanding (EX)

This category refers to any effort of extending, clarifying, or emphasizing one's own or another's individual or shared perception about the issue at hand. While AC/DC moves capture a speaker's attitude towards another's viewpoint, Expanding (EX) moves identify attempts to understand and interact with another's position or make one's own position more acceptable or understandable to the audience. Examples of such elaboration are the following (see also Hennessy et al., 2016): (a) contributions to the dialogue that build on, give examples, add to, reformulate or clarify one's own or others' contributions; (b) contributions that add something either in terms of content or in the way ideas are expressed. The repetition of one's own or others' ideas is not Expanding (it would be an irrelevant Stating). Table 4 shows examples of Expanding moves.

TABLE 4
Expanding moves

Example	Explanation
(S3) *How can you RECOGNIZE that [how the character feels]?* (IN)	S1 specifies and clarifies the question asked by S3. This echoic question is used to modify and specify what S3 is asking – for this reason it is an EX, and not an IN.
(S1) *No, how can you recognize that he feels like that?* (EX)	
(S1) *From his EXPRESSION.* (ST)	
(S4) *Yes, his expression, exactly. HIS ANNOYED EXPRESSION.* (EX)	Expanding (the student agrees with and specifies S1's viewpoint).
(S3) *That's THERE and there and there (points with the pen at different points on the piece of paper).* (ST)	
(S3) *His expression there at the bottom.* (EX)	S3 clarifies his own viewpoint expressed in the previous move.

4.1.2.2. Metadialogical (MD)

Metadialogical actions "describe the behavior of the speaker when he [she] is doing something else besides 'taking his [sic] turn'," not moving the conversation further but rather making a further contribution possible, relevant, and coherent (Labov and Fanshel, 1977, p. 60). Meta dialogical means talking about another move, turn, or discussion, in order to focus on a specific detail, which can be linguistic (prototypical case) or related to the subject matter (further focusing). A first case of the Metadialogical category refers to any verbal effort to explicitly make a connection between the current state of the dialogue (and/or the way it is understood) and its supposed/expected goal related to the activity under way. We call this *pragmatic metadialogical* type. An example of this type is meta-discourse about dialogue "ground rules" (Bereiter and Scardamalia, 2016; Littleton and Mercer, 2013). A second case concerns only the meaning of linguistic elements. This type of *linguistic metadialogical* moves can be: (a) requests of meaning explanation ("what does *x* mean?"); (b) requests of confirmation of understanding ("is my interpretation correct?"); (c) statements of lack of understanding ("I do not understand *x*;" "For me, *x* is *y*"); or (d) explanations of meaning ("*x* means *y*"). Examples of both types of Metadialogical (MD) moves are shown on Table 6.

TABLE 6
Moves coded as Metadialogical

Example	**Explanation**
We are so different that we cannot arrive at a common interpretation.	Pragmatic Metadialogical
I have already said it, a house is a building made for living or live together and a home is when you give a sentimental sense, emotional, which is yours or your family's.	Linguistic Metadialogical
(S1) *A home is where the heart is.*	Metadialogical discussion on the concept of home
(S4) *A house is a structure that {respects us}*	
(S1) *But they are talking in the sense of* Home. *You see it in English.*	
(S1) *Here in Portugal there is no kind of …* Home …	
(S1) *No, house is where you live!*	

4.1.2.3. Reasoning (RE)

This category refers to a class of conversational actions characterized by the disputable nature of the subject matter (Labov and Fanshel, 1977, p. 62), and includes arguments or counterarguments (where the doubt or potential dissent is taken for granted in the need of providing a justification). This code refers to any expression of a more or less justified idea about an issue at hand, which moves the dialogue forward. It includes the following cases (see also Hennessy et al., 2016): (a) explicitly acknowledging a shift of position *by providing a justification* (otherwise it would be Stating); (b) challenging others' arguments, beliefs or assumptions *by providing reasons* (otherwise it would be Accepting/Discarding); (c) *synthesizing or bringing together ideas*, or generalizing – when aimed at supporting a specific perspective; or (d) *making reasoning explicit* by using explanations, justifications, argumentation (providing an argument or a counterargument), analogies, or evidence, or formulating justified hypotheses. Examples of Reasoning (RE) moves are shown in Table 7.

TABLE 7
Reasoning moves

Example	Explanation
(S1) *Because I can be the same culture as her but maybe she is a man and I am a woman, and this already makes us different, for sure.*	A viewpoint (implicit in this case) is supported by a reason, an argument.
(S3) *So Pedro says it is through education that we learn how to be tolerant, I say it is through doing voluntary stuff, so what about writing "learn about volunteering"?*	A summary of the different positions is given to show the common aspects, addressing a difference by solving it.

4.1.2.4. Metadialogical Reasoning (MD_RE)

This type of move captures a unique combination of two types of moves, Metadialogical and Reasoning, and represents the highest level of potential transactivity. It refers to attacks to viewpoints or arguments based on the meaning of the viewpoint or the argument or the implicit premise that is taken for granted. An example illustrating this "reinforced" Metadialogical (MD_RE) code is presented in Table 8.

TABLE 8
Metadialogical reasoning moves

Example	Explanation
(S1) *Yes. [...] Hmm [...] I don't know... [home] It's also a place where you're supposed to rest... I mean-* (MD)	Metadialogical: Provides a definition of home
(S3) *So because I can rest, I can sleep, but I can sleep anywhere like that... you don't need a house!* (MD_RE)	MD_RE as the target of the attack is a definition
(S1) *It is, isn't it! It's just like, imagine... even if you- there are a lot of houses where you're at home, but you don't feel relaxed because your parents can be more severe or have big discussions. But because of that, it doesn't mean you don't feel at home, 'you see? So, this is super ... [...]* (MD_RE)	MD_RE as the target of the attacks is a definition and the student undermines it based on linguistic evidence.
(S5) *And when they say: oh, I don't want to go home because my parents argue. No, you don't want to go home, you don't want to be near your parents!* (MD_RE)	
(S1)*{unclear}[the baboon,] [the character of the story that is depicted on the Moon] is from Earth because he was different from other baboons. I think it's a metaphor for what we do.*	MD_RE as the student defends a symbolic interpretation

4.1.3. Across the levels of dialogicity: Inviting (IN)

This category potentially manifests all the other moves, as an IN move can be performed to elicit – inter alia – the interlocutor's expression of a viewpoint, agreement, argument, or definition of a concept. In classroom settings, teachers' Inviting moves can be also procedural: for example, the classical recitation questions used for testing students' recall are closer to the management of testing process than the request of a viewpoint (Alexander, 2018). The dialogicity of IN moves depends on the role of the speaker: teachers' questions requesting the interlocutors' viewpoints can be a method for involving the students by collecting their positions on an issue (cumulative talk). However, when such moves are authentically used for exploring the interlocutors' ideas (for instance, when performed in students' group discussions), they can be strong indicators of other-orientedness.

Typical cases of IN moves in small-group discussions can emerge in the following circumstances: (a) when a student invites other students to express their viewpoint on a certain topic, either by repeating a teacher's invitation or by genuinely "reaching out" to the other's point of view; (b) when a student invites other students to advance their own viewpoint on a certain interpretation, either by asking simply a request for confirmation, agreement or disagreement, or by inviting in a more elaborated way others' ideas – opening up the space of debate among the group. Table 5 presents examples of Inviting moves.

TABLE 5
Inviting moves

Example	Explanation
(T) *What do you think about the story's character? Is he nice?*	Inviting (the teacher is requesting the students' viewpoint)
(S2) *Yeah, OK, but right now what are you going to say about what are the father's expectations of the son?*	Inviting (the student is genuinely requesting another student's opinion)

4.1.4. Relevance

The degree of Relevance (low or high) is a distinct dimension of a move, which refers to how related a move is to the rest of the dialogue. In case of low-dialogic moves (Stating, Managerial, Accepting/Discarding), relevance captures the degree to which such moves are related to the topic under discussion or to the task/activity at hand. High-dialogic moves (Inviting, Expanding, Reasoning and Metadialogical) are classified as highly relevant when they refer to or address the other's move(s), continuing the dialogue by taking into consideration the interlocutor's contribution. In both cases, the "reasoning by exclusion" rule applies, namely: if it not irrelevant or lowly relevant, then it is highly relevant. The passage from a textual to a dialogic level (Macagno, 2019) is decided following this rationale:

1. Expanding. If a move expands the viewpoint proposed by the *same* speaker without considering the other moves that have occurred in the meanwhile, then it is Expanding with a low relevance (EX-).

2. Inviting. The level of relevance is low (IN-) when it is an invitation for someone to say what (s)he thinks, without a clear manifestation of the speaker's interest in better understanding the other's opinion or relation with the rest of the discourse. High relevance codes (IN+) usually refer to a previously stated contribution which needs to be further explained, clarified, justified, etc.
3. Reasoning. It is relevant (+) by default unless completely unrelated to the rest of the discourse (an argument is provided for the dialogic goal of addressing the interlocutor's actual or potential doubt). If an opinion is expressed without a reason, it is Stating.
4. Metadialogical. It is relevant when it addresses the previous move. When the MD move refers to the dialogue process or activity itself without any connection with the moves performed previously, then it is irrelevant (-). When a MD move refers to the dialogue process without the intention of a genuine reflection on the dialogue goals, then it is irrelevant (-).

Table 9 presents a coded excerpt example including the Relevance code (+/-).

TABLE 9
Excerpt coded with Relevance degree codes

Line	Speaker	Transcription (translated from PT)	Code	Relev.
1	T2	You said something quite interesting a while ago, you identified Aladdin…	IN	+
2	S2	ALADDIN, DUMBO, RED RIDING HOOD {unclear}	EX	+
3	T2	In the middle of all this diversity, are there any things in common? Where does Aladdin story come from? Is it from Europe? No?	IN	+
4	S3	It is from the Arabia	ST	+
5	T2	But it forms part… of the children's tales of the whole world, isn't it curious?	IN	+
6	S3	This one here is Romeo and Juliette! (laughs)	ST	-
7	S2	NO, THIS IS JUMANDJI!	DC	-
8		{Off task}		
9	T2	Do not only… {unclear}		
10	S5	I don't know, this is a quite strange scene, when I see this scene it reminds me of the Asians, but when I see the ox, it reminds me of Egypt. I don't know […] I don't know why	ST	+
11	S2	But the ox is up there. I don't know, I think there are more than one thing on the, on the same page	ST	+
12		{off task}		
13	S5	I REALLY think that this part here is the most recent and that it later passes on to […] on something older	ST	+
14	S1	I think NO-	DC	+
15	S2	BUT WATCH what the teacher SAID	MA	-

4.2. Degrees of dialogicity

As mentioned above, the categories can be broadly distinguished into lower and higher dialogic moves. Managerial (MA), Stating (ST), and Accepting/Discarding (AC/DC) are not prototypically used for understanding the other's perspective. Managerial moves do not involve an exchange of viewpoints nor understanding of others' ideas. Stating consists in advancing a viewpoint and thus it does not presuppose any interest in the other's viewpoint or position. Accepting/Discarding expresses the positioning of the speaker vis-à-vis the other's viewpoint, and similarly it does not require the consideration of another's perspective.

In contrast Expanding (EX), Inviting (IN), Metadialogical (MD), Reasoning (RE), and Metadialogical Reasoning (MD_RE) are necessarily dialogic, as they need to include, address, or consider the interlocutor's perspective. Expanding builds on another's contribution, specifying, describing, or developing it, or provides more details to the interlocutor to understand one own's viewpoint. Metadialogical moves are aimed at defining the meaning of the words used, namely establishing the common ground between the interlocutors, thus addressing sources of possible disagreements. The performance of Reasoning moves prototypically requires dialogic empathy, since to persuade someone the speaker needs to start from the premises that are presumed to be accepted by the interlocutor (Gilbert, 1997; Walton, 1992b). Metadialogical Reasoning is defined by the negotiation of common ground: the speaker provides reasons for a specific meaning or definition that constitutes the presupposition for understanding or coming to an agreement with the interlocutors.

The Inviting moves are more complex to classify, as they essentially require the other's reply as a condition for their successful performance, but they do not necessarily build on the other's viewpoint or request the expression thereof. The Inviting move can manifest all the other categories – from the procedural "testing" of a student's knowledge to the elicitation of the interlocutors' reasons underlying a position. The peculiar dialogical status of this move depends also on the speaker's role: while teachers' questions are not necessarily dialogical, students' inviting moves are normally considered as strong indicators of dialogicity, as they are genuine manifestations of elicitation of others' viewpoints (see for instance Teo, 2019)[3].

[3] Echoic uses of teachers' inviting questions are considered as Managerial moves.

4.3. Code predominance

The coding unit is the turn; however, a turn can express a plurality of moves (Macagno and Bigi, 2017). Thus, the principle of code predominance is essential, being used to decide how to code a turn when two or more distinct moves are expressed. The principle reflects the fact that a speaker is presumed to uptake the interlocutor's move and continue the dialogue that has been *proposed* thus far (Ducrot, 1972). Thus, the more dialogic code prevails over the less dialogic. An example is the following turn:

> *Yeah, you are right, but I think that the problem of migration needs to be considered as an international problem.*

The prevalence of the dialogic over the non-dialogic category is based on the assumption that sequences express one interactional (social) goal, namely one specific function that they play within the discourse (Macagno and Bigi, 2017; Merin, 1994, p. 238; Stubbs, 1983, chap. 8.2; Walton, 2007; Widdowson, 1979, p. 144). The core of the "social" or interactive act is the way that it *actually* modifies the interaction, namely the readjustments of the mutual communicative intents that it generates (Widdowson, 1979, p. 144).

In the example above, two distinct codes capture two distinct communicative intentions: an interactive one (an acknowledgment) and a dialogic one (an argument against a possible different viewpoint). The two intentions are not on the same level. The overall effect of the turn is to advance a grounded viewpoint, resulting in a deeper "readjustment" of the interlocutor's communicative options (Ducrot, 1972). The first intention is ancillary to the latter, acting as a cohesion mechanism. Therefore, the above utterance will be coded as a Reasoning (RE) rather than an Accepting (AC) move.

The operationalization of dialogue other-orientedness proposed in this section can become a methodological tool only if validated. In the following section, the validation procedure and results will be presented, showing how the categories advanced can be used as an instrument for dialogue analysis.

5. Validation of the coding scheme

The passage from the operationalization of a concept through coding categories to a coding scheme that is usable for authentic analytical purposes, is guaranteed by its validity and reliability. A very basic criterion for the *validity* of a coding scheme is its capability to orient coders towards the theoretical grounding that it is intended to re fer to (Poole and Folger, 1981). Two essential aspects of validity are transparency and relationship between the codes and the underlying concepts (Potter and Levine-Donnerstein, 1999). Theory-driven approaches that deconstruct existing theories into codes need to be based on a standard "correct coding" procedure, which should guide coders. Complying with and integrating expert feedback into the validation process can add incremental value to the scheme by strengthening its ties to the underlying constructs. Thus, the involvement of additional expertise can provide evidence for drawing inferences on the validity of a coding scheme (Kane, 2006; see also Cook et al., 2016). *Reliability* is conceptualized as the consistency of a measure across multiple assessments or multiple ratings of the same event (Cook and Campbell, 1979). In the context of coding qualitative data, interrater reliability is the widely used term for the extent to which independent (or blind) coders evaluate a characteristic of a coding unit in the same way, that is, if all coders make comparable judgements, then the data is regarded as reliable (Krippendorff, 2004; Lombard et al., 2002). Consequently, high reliability of a scheme that has been well-grounded in theory would be the most important prerequisite to allow its wider use in other contexts and with other samples.

5.1. Sample

For the validation, the corpus described in Section 4 was used. Out of a total of 21 transcribed in-class discussions of the four countries, 31 excerpts ranging from 42 to 174 units (turns) were randomly chosen for the validation process in their original languages. To ensure a high degree of variety in the data sample proportion of whole class and small group discussions, the three age groups were balanced within countries (Table 10).

TABLE 10
Sample data

Country	Total number of units for coding	Number of units in Whole Class or Small Group discussions	Number of lines in age group 1 (pre-primary), 2 (primary) or 3 (secondary)
Experts´ feedback			
England	229	95 (WC) 134 (SG)	50 (1) 128 (2) 51 (3)
Pilot test (first phase)			
Germany	366	125 (WC) 239 (SG)	77 (1) 150 (2) 139 (3)
Portugal	199	117 (WC) 82 (SG)	74 (1) 42 (2) 83 (3)
Spain	345	164 (WC) 181 (SG)	68 (1) 103 (2) 174 (3)
Total (pilot)	910	406 (WC) 502 (SG)	219 (1) 295 (2) 396 (3)
Final test (second phase)			
Germany	330	200 (WC) 130 (SG)	87 (1) 76 (2) 167 (3)
Portugal	198	109 (WC) 89 (SG)	40 (1) 93 (2) 65 (3)
Spain	198	124 (WC) 74 (SG)	67 (1) 69 (2) 62 (3)
Total (final)	726	433 (WC) 293 (SG)	194 (1) 238 (2) 294 (3)
Total (overall)	**1363**	**840 (WC)** **796 (SG)**	**413 (1)** **533 (2)** **690(3)**

5.2. Development of the codebook

A first version of the codebook was subjected to expert validation. It was given to two experts in Education, Dialogue, and Argumentation who provided detailed feedback on its appropriateness and usability. After blindly coding a randomly chosen sample of 229 units of the English corpus (Table 10), they answered a list of three categories of questions concerning the codebook: content-related issues (e.g., description of the categories), functional aspects (e.g., clear assignment to the categories), and general recommendations. The feedback focused on the use and coherence of the terminology, the sufficiency and exhaustiveness of the coding categories, the description of the coding categories, the dimensions of relevance and other-orientedness, the disambiguation of coding categories, and the appropriateness of the segmentation rules. The codebook was then given to six blind and untrained coders from Germany, Portugal, and Spain. The coding of a sample size ranging between 199 and 366 units across countries, separate from the sample used for final reliability tests, was done independently and without any consultation or guidance. Based on the expert and the coder feedback, the codebook was refined before conducting the final reliability tests, shown on Table 11.

TABLE 11
Expert and coder feedback

Scopes	Expert Feedback	Coder Feedback	Addressed in Codebook
Use of terms	Distinction between Relevance and Transactivity Explanation of Dialogical Transactivity Reduction of technical terms	Distinction of Relevance and Transactivity	Degree of Relevance is manifested Dialogical Transactivity is defined as low-dialogical and high-dialogical moves More explicit explanation, less technical terms
Content and elements	Differences between code categories (especially for RE and EX, for ST, for IN and for MD) Demand for more examples and explanations Defining AC/DC as either one or two categories	Further distinction between code categories (especially for RE and EX, for ST and for IN) Demand for more examples for each code categories	Distinctions between each code categories in additional subchapters Explained examples for each code category AC/DC as one category
Dimensions of Relevance and Transactivity	Defining differences and level of transactivity between code categories Distinction between high and low transactivity Transactivity rules for specific cases (IN, MA and revoicing utterances)	Description for the assignment of high or low transactivity Importance of dialogical moves Transactivity for IN as a specific case	Defining 2 dimensions of Relevance: low-dialogical and high-dialogical moves Definition of Dialogicity Criterion and Clarity Criterion
Content-related rules for specific codes (Disambiguation of codes)	Explicit rules and guidelines for prevalence of codes (in cases if more than one code corresponds to one utterance)	Missing criteria of prevalence between two or more codes	Definition of code predominance with subchapters “Dialogicity criterion” and “Clarity criterion”
Disambiguation of utterances	Demand for rules for special cases like unintelligible utterances, not assignable codes, interruptions, overlapping utterances	Demand for rules for special cases like interruptions, unfinished turns, nonverbal communication, off-topic utterances, repetitions, summarizing arguments, requests of clarification	Additional information in extra chapters for “Not coded moves”, “Clarity criterion” and “Moves continued in another turn”
Transcript	Demand for brief description at the beginning of each transcript segment of general content	Demand for description of linguistic form of utterances	
Recommendations	Training periods for coders Notion system for script		

To determine the valid reliability indices, the final version of the codebook was tested in a series of interrater tests in the different countries, based on the coding of a new randomly chosen representative sample out of the full data set (Table 10) (Neuendorf, 2002; see Lacy and Riffe, 1996). To prevent coder drift during coding (Wolfe et al., 2001), agreement was checked informally and only between coders after coding the first 30 units in each team. Interrater agreement in the final test between and across countries was moderate to good as indicated by the reliability indices on the categorical level (Table 12).

TABLE 12
Interrater reliabilities in pilot and final tests

Krippendorff's Alpha (categorial)		
	Pilot Test	**Final Test**
Portugal	.36	.67
Spain	.55	.68
Germany	.41	.85
All countries	.46	.77

Detailed interrater agreement results for each category as obtained in the final test are presented in Table 13, with overall interrater agreement of a = .77, normally assessed as fully satisfactory (Carletta et al., 1997).

TABLE 13
Detailed interrater reliability results in final test per category

	PORTUGAL			**SPAIN**			**GERMANY**			**All 3 Countries**		
	Agree ment	**κ**	**α**	**Agree ment**	**κ**	**α**	**Agree ment**	**κ**	**α**	**Agree ment**	**κ**	**α**
Not Coded	.96	.85	.85	.96	.65	.65	.99	.95	.95	.97	.88	.88
MA	.77	.52	.52	.96	.57	.57	.97	.82	.82	.91	.67	.67
ST	.83	.24	.24	.81	.50	.50	.94	.8	.8	.87	.59	.59
AC/DC	.92	.51	.51	.95	.78	.78	.98	.92	.92	.96	.80	.80
EX	.92	.37	.37	.91	.51	.52	.97	.84	.84	.94	.61	.61
IN	.91	.59	.59	.90	.75	.75	.95	.85	.85	.92	.77	.77
MD	.97	.27	.28	.99	.50	.50	.97	.71	.71	.98	.63	.63
RE	.92	.35	.35	.95	.66	.66	.97	.82	.82	.95	.68	.68
MD_RE	1	-	-	1	-	-	1	1	1	1	1	1
Relevance	.88	.66	66	.96	.75	0.75	.99	.99	.98	.96	.7	.7

κ: Cohen's Kappa
α: Krippendorff's Alpha

Two elements need to be considered in the interpretation of this table. First, the Portuguese Alpha values are much lower than the Spanish and especially German ones. These differences can be explained considering the different training and coding conditions of the coders. The German and Spanish coders were trained in discourse and data analysis, and they were Ph.D. students or post-doc researchers. In contrast, the Portuguese coders were Master students with little training in this field of research. Moreover, the teams had different acquaintance with the codebook and different possibility of reviewing their own coding. While the German and Spanish coders were exclusively allocated studying and applying the codebook, the Portuguese team was allocated also to other different work packages, which resulted in lower familiarity with the codes.

A second aspect of this reliability analysis concerns the frequency of the categories. Some categories have very low frequencies (on average for all three countries, MD= 2.5%; RE=3,5%; MD_RE = 1,5%) while others very high (on average, MA=17%; ST=26%; AC/DC=11%; EX=5,5%; IN=18,5%). No MD RE codes were found in the Portuguese and Spanish coded sample; therefore, the Cohen's Kappa and the Krippendorff's Alpha values cannot be reported. Moreover, the frequency of these codes across the different countries' datasets varied noticeably, especially for the categories RE (Spanish RE moves were almost 4 times more frequent than the Portuguese ones, and twice as frequent as the German ones) ST (Spanish ST moves were 3.5 times more frequent than the German ones, and twice as frequent as the Portuguese ones), and IN (Spanish IN moves were 3.7 times more frequent than the Portuguese ones, and twice as frequent as the German ones).

6. Other-orientedness, common ground, and empathy

The distinction between degrees of dialogicity is intended to identify when and how a discussion unveils potential or actual differences in what is commonly taken for granted. A premise, or more generally a proposition, is treated as commonly accepted when it can be considered as part of the cultural background, i.e., what people in a community are disposed to accept (Ducrot, 1972, 1966). However, deeper forms of disagreement lie in these unstated assumptions (or presuppositions), resulting in mismatching – or uncommon – "grounds." Higher dialogicity codes can be used for identifying sequences in which the ground between the interlocutors is more likely to be brought to light, addressed, and negotiated. In particular, Reasoning moves are grounded on implicit generalizations (Walton, 2008) that can mirror cultural differences, while Metadialogical moves bring to light one of the deepest aspects of the core common ground (Kecskes and Zhang, 2013, 2009), i.e., the meaning of the words and concepts used. Metadialogical Reasoning captures the negotiations of this core common ground, providing reasons (and thus an opportunity) for developing a shared cultural basis for the discussion. The frequency of these moves can be a sign of the presence of negotiations of the interlocutors' grounds (and thereby the other-oriented nature of the interaction). To illustrate the relationship between other-orientedness and the development of common ground, some excerpts from our corpus are presented.

The first excerpt represents an example of unaddressed disagreements that could reveal deeper differences in common ground. The students were instructed to discuss the interpretation of a short movie (*Papa's Boy*) whose main character is a little mouse who wants to be a ballerina, contrary to its father's desire The students advance arguments (1, 5, and 9) that show a clear difference of interpretation (the little mouse is interpreted as a female vs. male due to its interests) which is, however, left unaddressed (the teacher encourages other contributions through managerial moves). The higher-dialogical moves reveal an aspect of the common ground that is potentially controversial (differences between males and females), as shown in 5 (S7 refers to the mouse as a he-mouse)[4].

[4] Discussions presented in Case 1, Case 2, and Case 4 are stimulated by the wordless short-animated film "Papa's boy" (original title in Finnish: Isän poika) by Leevi Lemmetty (2010) (available here: https://www.youtube.com/watch?v=vTmUpQbJ_HI). For more information about the pedagogical materials used as part of our project, see the project's website: https://dialls2020.eu/

CASE 1
Dialogicity and apparent common ground

Line	Speaker	Transcription	Code
1.	S5	A difference is that the little mouse is a girl, and her dad is a boy and it's normal not to have the same taste, that's it.	RE
2.	T	Right. Is it natural? Yes.	MA
3.	S6	I noticed that the dad was practicing boxing and that the eerr baby could fly.	ST
4.	T	Uh-huh. Yes.	MA
5.	S7	The little mouse could fly because he knew ballet and the dad didn't know ballet.	RE
6.	T	Hmmm. Indeed! Well. Let's continue. Yes, S8?	MA
7.	S8	Err then the first time she danced ballet, her dad realized that he thought he should teach her boxing like kickboxing, I remember.	ST
8.	T	Uh-huh. That's what he thought.	AC
9.	S8	And she didn't like because she was not, girls don't learn boxing easily	RE

Case 2 (concerning the interpretation of the short movie described above) presents a different dialogic pattern. At 1, the teacher invites S1 to provide the reasons underlying his viewpoint. The student replies with an argument presupposing an implicit premise (the fact that the dad was a boxer causes his disappointment at seeing the son dancing), which S1 mistakenly takes as commonly accepted. S2 reconstructs this premise and starts a dialogue on its reasonableness.

CASE 2
Dialogicity and the explanation of common ground

Line	Speaker	Transcription	Code
1.	T	Right. The dad felt weird at the start because hi his boy danced. BUT WHY?	IN
2.	S1	Um, uh, because he the dad was a boxer.	RE
3.	S2	And if the dad is a boxer the boy isn't allowed to dance?	MD
4.	S1	Nope.	ST
5.	T	Why not?	IN
6.	S1	No, he is allowed to dance.	DC
7.	S4	Because he should actually be a boxer too.	RE
8.	S1	The dad wanted the son to be a boxer.	EX

The dialogic scenario is more complex in Case 3, which addresses the differences between people, and the relationship between climate and culture. S4's argument is followed by several reasoning moves focused on the causes of seasons (3 and 4) and more importantly cultural habits and seasons (8 and 9). Finally, at 11, S7 undermines the arguments provided, classifying the causes of different traditions as cultural.

Line	Speaker	Transcription	Code
1.	S7	[Seasons?] (S7 to S4)	IN
2.	S4	Yeah, it makes a huge difference. Listen. I was in the Caribbean and uh was with my family there and the mum there asked me, hey do they really have four seasons? And I just- {} at first, I thought what a {weird} question. Then I realized, of course- they only have monsoon season and, and, and dry season and not four seasons. That has a big influence on the culture too.	RE
3.	S7	But (confusing talk on the connection between weather and seasons)	RE
4.	S4	Nah, definitely not. The number of seasons you have isn't a question of weather. {It's a question} of climate.	RE
5.	S3	But in everyday life now, what differences do we have there?	IN
6.	S4	[...] but Christmas the way we celebrate it or Advent and that kind of thing, that, that, that cozy idea of winter and snow and so on, it can't for example- Where are you from? (S4 talks to S5)	IN
7.	S5	{Iraq.}	ST
8.	S4	Can't exist in Iraq for example, because they have completely different climatic conditions because they don't have four seasons. (S4 talks to S7)	RE
9.	S4	If you tried to take the tradition to Iraq, without the Christian tradition but just like, if you try to bring that Coca-Cola Christmas to Iraq somehow, it won't work because you don't have any snow- [they've never seen snow.] Christmas trees don't mean anything to them, Christmas trees covered in snow don't mean anything to them because they're missing the cultural reference. That's why seasons really are very important.	RE
10.		[No winter, no nothing]	EX
11.	S7	{For me that would be culture- not the season}	MD
12.	S4	Yeah, [yeah] but the season, it, it, it shapes the culture. It's all connected.	MD

Case 4 is more complex from a dialogic point of view. The students are discussing the interpretation of the short movie *Papa's Boy*, and in particular the sex of the dancing mouse. S1 advances an argument based on three implicit premises: a) that dancing is a girl's sport; b) that dancing means behaving like a girl; and 3) that it is inacceptable that a boy behaves like a girl. This reasoning move leads to a metadialogical rebuttal (Line 2), in which S8 rejects the premises taken for granted. S1 replies by making his tacit premise explicit, distinguishing the possibility of doing whatever one wants from its appropriateness. The last move is again a metadialogical attack against the first move, in which the identification between gender and behavior is challenged.

CASE 4
Dialogicity and the negotiation of common ground

Line	Speaker	Transcription	Code
1.	S1	I think the mouse like [..] can't all the time deal with girly stuff because he has like a whole world ahead of him and he's supposed to like deal with for example basketball, football, {boys'} [stuff] [...] and he can't just be a girl all the time [...]	RE
2.	S8	I think thaaaat... that why you say is wrong because [..] any boy can do whatever a girl can do and[..] any girl can play whatever a boy does, and any girl can play football and any boy can play, uh, dunno with dolls or such things. Because it's about what everyone loves, whatever they love they'll do and persist with it. Like in the, in class, we have it in class we always play ball. Girls and boys as well, and you can see it just in front of your eyes.	MD
3.	S1	Yes, I know but, I didn't mean that. I meant, he can play whatever he feels like but, like, he ca-, he can do whatever he feels like but, like-	MD
4.	S28	I actually differ from you in opinion S1 because, for example, for example there are boys who actually like, actually the boys in class have like always with ball like if they forget what was last time, dodgeball or football, so there are some that the boys are saying football and the girls are saying dodgeball. So, there are also boys. For instance, like S9. Like S9 he sometimes also says dodgeball. And also, for instance in my summer camp, can't remember for example a year or two years ago for instance there was a girl who really loved playing football there really {unclear}	MD

These examples illustrate the relationship between other-orientedness and common ground. Communication is based on what is commonly referred to as the "common ground," namely assumptions that need to be shared by the interlocutors. However, often our projection of what is uncontroversial overshadows actual cultural differences (Mustajoki, 2012). Other-oriented moves allow the ground that is presumed to be common to be brought to light, discussed, and negotiated. The awareness of the relationship between the levels of dialogicity and the understanding of another's perspective on what is culturally or communally shared can offer the possibility of creating a ground that is truly common. This common basis of experience is the requirement not only for communicating effectively across different cultures, but also for understanding the others and experiencing their emotions.

7. Conclusion

The role of empathy in communication, pragmatics, and rhetoric has been acknowledged as crucial, but its analysis has been almost neglected. If we look at the field of pragmatics, we can see how politeness is grounded on an empathetic attitude towards the interlocutor (Fukushima and Haugh, 2014), while in intercultural communication empathy is identified as a necessary element and even a pre-requisite (Teal and Street, 2009). In argumentation, understanding the other's commitments is a requirement for developing arguments that can be acceptable (Gilbert, 1997; Walton and Krabbe, 1995). In rhetoric, adapting the discourse to the audience – namely, making it suited to the common ground of the hearers – is the role of persuasion. Despite its crucial importance, empathy has been little investigated as a pragmatic and communicative phenomenon, leaving the problem of its textual detection almost entirely open.

This paper has addressed the challenge of operationalizing dialogic empathy from a pragmatic and discursive perspective by proposing and successfully validating a coding scheme capturing other-orientedness. We showed how the degrees of dialogic empathy can be captured by considering the pragmatic purpose of the move: some moves necessarily involve the other or the other's perspective, while others are purely monological. However, the structure of a move is only one dimension of other-orientedness: even a potentially high dialogic move does not manifest any dialogic empathy if it is unconnected with the interlocutors' talk. Thus, a highly dialogical move can be actually other-oriented only when it is used relevantly in a dialogue. These two coding criteria –

dialogicity and relevance – capture the conditions of deep understanding, identifying when the other's viewpoint is acknowledged and addressed in its entirety, including the reasons underlying it.

References

Alexander, R., 2018. Developing dialogic teaching: genesis, process, trial. *Res. Pap. Educ.* 33, 561–598. https://doi.org/10.1080/02671522.2018.1481140

Aristotle, 2010. *Rhetoric.* Cambridge University Press, Cambridge, UK.

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

Bach, K., Harnish, R., 1979. *Linguistic communication and speech acts.* MIT Press, Cambridge, MA.

Barrett, M. (Ed.), 2013. *Interculturalism and multiculturalism: similarities and differences.* Council of Europe, Strasbourg, France, France.

Battaly, H., 2011. Is empathy a virtue?, in: Coplan, A., Goldie, P. (Eds.), *Empathy: Philosophical and Psychological Perspectives.* Oxford University Press, Oxford, UK, pp. 277–301.

Ben-Ze'ev, A., 2000. *The subtlety of emotions.* MIT Press, Cambridge, MA.

Bereiter, C., Scardamalia, M., 2016. "Good Moves" in knowledge-creating dialogue. *Qwerty-Open Interdiscip. J. Technol. Cult. Educ.* 11, 12–26.

Berkowitz, M., Gibbs, J., 1985. The process of moral conflict resolution and moral development. *New Dir. Child Adolesc. Dev.* 1985, 71–84. https://doi.org/10.1002/cd.23219852907

Berkowitz, M., Gibbs, J., 1983. Measuring the developmental features of moral discussion. *Merrill-Palmer Q.* 399–410.

Boyd, M., Markarian, W., 2011. Dialogic teaching: talk in service of a dialogic stance. *Lang. Educ.* 25, 515–534. https://doi.org/10.1080/09500782.2011.597861

Braet, A., 1999. The enthymeme in Aristotle's Rhetoric: From argumentation theory to logic. *Informal Log.* 19, 101–117. https://doi.org/10.22329/il.v19i2.2322

Buber, M., 2002. *Between man and man.* Routledge, New York, NY.

Buber, M., 1999. *Pointing the way: Collected essays.* Humanity Books, London, UK.

Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J.C., Anderson, A.H., 1997. The reliability of a dialogue structure coding scheme. *Comput. Linguist.* 23, 13–31.

Carter, M., 1988. Stasis and kairos: Principles of social construction in classical rhetoric. *Rhetor. Rev.* 7, 97–112.

Chin, C., 2006. Classroom interaction in science: Teacher questioning and feedback to students' responses. *Int. J. Sci. Educ.* 28, 1315–1346. https://doi.org/10.1080/09500690600621100

Christodoulou, A., Osborne, J., 2014. The science classroom as a site of epistemic talk: A case study of a teacher's attempts to teach science based on argument. *J. Res. Sci. Teach.* 51, 1275–1300. https://doi.org/10.1002/tea.21166

Clark, H., 1996. *Using language*. Cambridge University Press, Cambridge, UK.

Clarke, S., Resnick, L., Rosé, C.P., 2015. Dialogic instruction: A new frontier, in: Corno, L., Anderman, E. (Eds.), *Handbook of Educational Psychology*. Routledge, New York, NY, pp. 392–403.

CoE (Council of Europe), 2018. *Reference framework of competences for democratic culture*. Strasbourg, France.

Cook, D.A., Kuper, A., Hatala, R., Ginsburg, S., 2016. When assessment data are words: validity evidence for qualitative educational assessments. *Acad. Med.* 91, 1359–1369. https://doi.org/10.1097/ACM.0000000000001175.

Cook, T., Campbell, D., 1979. *Quasi-experimentation: Design and analysis for field settings*. Rand-McNally, New York, NY, NY.

Dawson, V.M., Venville, G., 2010. Teaching strategies for developing students' argumentation skills about socioscientific issues in high school genetics. *Res. Sci. Educ.* 40, 133–148. https://doi.org/10.1007/s11165-008-9104-y

Ducrot, O., 1972. *Dire et ne pas dire*. Hermann, Paris, France.

Ducrot, O., 1966. "Le roi de France est sage". Implication logique et présupposition linguistique. *Etudes Linguist. appliqu*ée 4, 39–47.

Elliott, R., Bohart, A., Watson, J., Greenberg, L., 2011. Empathy. *Psychotherapy* 48, 43–49. https://doi.org/10.1037/a0022187

Erduran, S., Simon, S., Osborne, J., 2004. TAPping into argumentation: Developments in the application of Toulmin's Argument Pattern for studying science discourse. *Sci. Educ.* 88, 915–933. https://doi.org/10.1002/sce.20012

Felton, M., Kuhn, D., 2001. The development of argumentive discourse skill. *Discourse Process.* 32, 135–153. https://doi.org/10.1080/0163853X.2001.9651595

Ford, M., 2008. Disciplinary authority and accountability in scientific practice and learning. *Sci. Educ.* 92, 404–423. https://doi.org/10.1002/sce.20263

Fukushima, S., Haugh, M., 2014. The role of emic understandings in theorizing im/politeness: The metapragmatics of attentiveness, empathy and anticipatory inference in Japanese and Chinese. *J. Pragmat.* 74, 165–179. https://doi.org/10.1016/j.pragma.2014.08.004

Gehlbach, H., 2004. Social perspective taking: A facilitating aptitude for conflict resolution, historical empathy, and social studies achievement. *Theory Res. Soc. Educ.* 32, 39–55. https://doi.org/10.1080/00933104.2004.10473242

Gibbs, R., 1987. Mutual knowledge and the psychology of conversational inference. *J. Pragmat.* 11, 561–588. https://doi.org/10.1016/0378-2166(87)90180-9

Gilbert, M., 1997. *Coalescent argumentation, Argumentation*. Lawrence Erlbaum Associates, Mahwah, NJ.

Goldie, P., 2000. *The emotions: A philosophical exploration*. Oxford University Press, Oxford, UK.

Goldman, A., 2006. *Simulating minds: The philosophy, psychology, and neuroscience of mindreading.* Oxford University Press, New York, NY.

Gough, J., Tindale, C., 1985. "Hidden" or "missing" premises. *Informal Log.* 7, 99. https://doi.org/10.22329/il.v7i2.2708

Grice, P., 1957. Meaning. *Philos. Rev.* 66, 377–388. https://doi.org/10.2307/2182440

Hamblin, C.L., 1970. *Fallacies*. Methuen, London, UK.

Hammond, S., Anderson, R., Cissna, K., 2003. The problematics of dialogue and power. *Ann. Int. Commun. Assoc.* 27, 125–157. https://doi.org/10.1080/23808985.2003.11679024

Hennessy, S., Rojas-Drummond, S., Higham, R., Márquez, A.M., Maine, F., Ríos, R.M., García-Carrión, R., Torreblanca, O., Barrera, M.J., 2016. Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learn. Cult. Soc. Interact.* 9, 16–44. https://doi.org/10.1016/j.lcsi.2015.12.001

Hirsch, E., 1983. Cultural literacy. *Am. Scholar* 52, 159–169.

Hitchcock, D., 1998. Does the traditional treatment of enthymemes rest on a mistake? *Argumentation* 12, 15–37. https://doi.org/10.1023/A:1007738519694

Howe, C., Hennessy, S., Mercer, N., Vrikki, M., Wheatley, L., 2019. Teacher-student dialogue during classroom teaching: Does it really impact on student outcomes? *J. Learn. Sci.* 28, 462–512. https://doi.org/10.1080/10508406.2019.1573730

Ickes, W., 1993. Empathic accuracy. *J. Pers.* 61, 587–610. https://doi.org/10.1111/j.1467-6494.1993.tb00783.x

Johnson, D., 1975. Cooperativeness and social perspective taking. *J. Pers. Soc. Psychol.* 31, 241–244. https://doi.org/10.1037/h0076285

Kane, M., 2006. Validation, in: Brennan, R. (Ed.), *Educational Measurement 4th Ed.* Praeger, Westport, CT, pp. 17–64.

Kecskes, I., Zhang, F., 2013. On the dynamic relations between common ground and presupposition, in: Capone, A., Lo Piparo, F., Carapezza, M. (Eds.), *Perspectives in Pragmatics, Philosophy & Psychology.* Springer, Cham, Switzerland, pp. 375–395.

Kecskes, I., Zhang, F., 2009. Activating, seeking, and creating common ground: A socio-cognitive approach. *Pragmat. Cogn.* 17, 331–355. https://doi.org/10.1075/pc.17.2.06kec

Kinneavy, J., Eskin, C., 2000. Kairos in Aristotle's rhetoric. *Writ. Commun.* 17, 432–444. https://doi.org/10.1177/0741088300017003005

Krippendorff, K., 2004. *Content analysis: An introduction to its methodology.* Sage, Thousand Oaks, CA.

Labov, W., Fanshel, D., 1977. *Therapeutic discourse: Psychotherapy as conversation.* Academic Press, New York, NY.

Lacy, S., Riffe, D., 1996. Sampling error and selecting intercoder reliability samples for nominal content categories. *Journal. Mass Commun. Q.* 73, 963–973.

Leech, G., 1983. *Principles of pragmatics*. Longman, London, UK.

Levinson, S., 1983. *Pragmatics*. Cambridge University Press, Cambridge, UK.

Lipman, M., 2003. *Thinking in education*. Cambridge University Press, New York, NY, NY.

Lipps, T., 1903. *Ästhetik: Psychologie des Schönen und der Kunst: Grundlegung der Ästhetik, Erster Teil*. Voss, Hamburg, Germany.

Littleton, K., Mercer, N., 2013. *Interthinking: Putting talk to work*. Routledge, Abingdon, UK.

Lombard, M., Snyder-Duch, J., Bracken, C.C., 2002. Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Hum. Commun. Res.* 28, 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Macagno, F., 2019. Coding relevance. *Learn. Cult. Soc. Interact.* 100349. https://doi.org/10.1016/j.lcsi.2019.100349

Macagno, F., 2018a. Evidence and presumptions for analyzing and detecting misunderstandings. *Pragmat. Cogn.* 24, 263–296. https://doi.org/10.1075/pc.17034.mac

Macagno, F., 2018b. A dialectical approach to presupposition. *Intercult. Pragmat.* 15, 291–313. https://doi.org/10.1515/ip-2018-0008

Macagno, F., Bigi, S., 2017. Analyzing the pragmatic structure of dialogues. *Discourse Stud.* 19, 148–168. https://doi.org/10.1177/1461445617691702

Maine, F., Cook, V., Lähdesmäki, T., 2019. Reconceptualizing cultural literacy as a dialogic practice. *London Rev. Educ.* 17, 383–392. https://doi.org/10.18546/LRE.17.3.12

Mercer, N., 2004. Sociocultural discourse analysis. *J. Appl. Linguist.* 1, 137–168.

Merin, A., 1994. Algebra of elementary social acts, in: Tsohatzidis, S. (Ed.), *Foundations of Speech Act Theory*. Routledge, London, UK, pp. 242–272.

Mustajoki, A., 2012. A speaker-oriented multidimensional approach to risks and causes of miscommunication. *Lang. Dialogue* 2, 216–243. https://doi.org/10.1075/ld.2.2.03mus

Neuendorf, K., 2002. *The content analysis guidebook*. Sage, Thousand Oaks, CA.

Nystrand, M., Gamoran, A., Kachur, R., Prendergast, C., 1997. *Opening Dialogue: Understanding the Dynamics of Language and Learning in the English Classroom*. Teachers College Press, New York, NY.

Poole, M.S., Folger, J.P., 1981. Modes of Observation and The Validation of Interaction Analysis Schemes. *Small Gr. Behav.* 12, 477–493. https://doi.org/10.1177/104649648101200406

Potter, W.J., Levine-Donnerstein, D., 1999. Rethinking validity and reliability in content analysis. *J. Appl. Commun. Res.* 27, 258–284. https://doi.org/10.1080/00909889909365539

Preston, S., de Waal, F., 2002. Empathy: Its ultimate and proximate bases. *Behav. Brain Sci.* 25, 1–75. https://doi.org/10.1017/S0140525X02000018

Rapanta, C., 2019. *Argumentation strategies in the classroom.* Vernon press, Wilmington, DE.

Rapanta, C., Cascalheira, D., Gil, B., Gonçalves, C., Garcia, D., Morais, R., Pereira, J.R., Čermáková, A., Maine, F., Peck, J., Brummernhenrich, B., Jucks, R., Petronytė, M., Valančienė, D., Juskiene, V., Badaukienė, R., Eigminienė, D., Zaleskienė, I., Garcia-Mila, M., Remesal, A., Castells, N., Gilabert, S., Miralda-Banda, A., Luna, J., Vrikki, M., Evagorou, M., Chatzianastasi, M., Karousiou, C., Papanastasiou, E., Stylianou-Georgiou, A., Rodosthenous, M., Talli, C., Cohen, I., Shalom Greenberg, C., Bar, N., Sarfati, N., Schwarz, B., 2020. *Dialogue and Argumentation for Cultural Literacy Learning in Schools: Multilingual Data Corpus.* Zenodo. https://doi.org/10.5281/ZENODO.4058183

Rapanta, C., Vrikki, M., Evagorou, M., 2021. Preparing culturally literate citizens through dialogue and argumentation: Rethinking citizenship education. *Curric. J.* 32, 475–494. https://doi.org/10.1002/curj.95

Rogers, C., 1980. *A way of being.* Houghton Mifflin Company, Boston, MA.

Sarangi, S., 2007. Other-orientation in patient-centred healthcare communication: Unveiled ideology or discoursal ecology, in: Garzone, G., Sarangi, S. (Eds.), *Discourse, Ideology and Specialized Communication. Special Issue of Linguistic Insights.* Peter Lang, Berlin, Germany, pp. 39–71.

Sarangi, S., 1998. "I Actually Turn My Back on [Some] Students": The Metacommunicative Role of Talk in Classroom Discourse. *Lang. Aware.* 7, 90–108. https://doi.org/10.1080/09658419808667103

Scheler, M., 2017. *The nature of sympathy.* Routledge, London, UK, and New York, NY.

Schiffer, S., 1972. *Meaning.* Oxford University Press, Oxford, UK.

Schmid, P., 2001. Comprehension: The art of not knowing. Dialogical and ethical perspectives on empathy as dialogue in personal and person-centred relationships, in: Haugh, S., Merry, T. (Eds.), *Empathy.* PCCS Books, Ross-on-Wye, UK, pp. 53–71.

Scott, P., Mortimer, E., Aguiar, O., 2006. The tension between authoritative and dialogic discourse: A fundamental characteristic of meaning making interactions in high school science lessons. *Sci. Educ.* 90, 605–631. https://doi.org/10.1002/sce.20131

Searle, J., 1975. Indirect speech acts, in: Cole, P., Morgan, J. (Eds.), *Syntax and Semantics Volume 3: Speech Acts.* Academic Press, New York, NY, pp. 59–82.

Shady, S.L.H., Larson, M., 2010. Tolerance, empathy, or inclusion? Insights from Martin Buber. *Educ. Theory* 60, 81–96. https://doi.org/10.1111/j.1741-5446.2010.00347.x

Sorensen, R., 1988. Are enthymemes arguments? *Notre Dame J. Form. Log.* 29, 155–159.

Sperber, D., Wilson, D., 1995. *Relevance: Communication and cognition.* Blackwell Publishing Ltd, Oxford, UK.

Stubbs, M., 1983. *Discourse analysis: The sociolinguistic analysis of natural language*. University of Chicago Press, Chicago, IL.

Sullivan, D.L., 1992. Kairos and the rhetoric of belief. *Q. J. Speech* 78, 317–332. https://doi.org/10.1080/00335639209383999

Teal, C., Street, R., 2009. Critical elements of culturally competent communication in the medical encounter: A review and model. *Soc. Sci. Med.* 68, 533–543. https://doi.org/10.1016/j.socscimed.2008.10.015

Teo, P., 2019. Teaching for the 21st century: A case for dialogic pedagogy. *Learn. Cult. Soc. Interact.* 21, 170–178. https://doi.org/10.1016/j.lcsi.2019.03.009

Tindale, C., 1999. *Acts of arguing: A rhetorical model of argument*. State University of New York Press, Albany, NY.

Untersteiner, M., 1954. *The sophists*. Basil Blackwel, Oxford, UK.

Verdonik, D., 2010. Between understanding and misunderstanding. *J. Pragmat.* 42, 1364–1379. https://doi.org/10.1016/j.pragma.2009.09.007

Viano, C.A., 1955. *La logica di Aristotele*. Taylor, Torino, Italy.

Vogel, F., Kollar, I., Ufer, S., Reichersdorfer, E., Reiss, K., Fischer, F., 2016. Developing argumentation skills in mathematics through computer-supported collaborative learning: the role of transactivity. *Instr. Sci.* 44, 477–500. https://doi.org/10.1007/s11251-016-9380-2

Voloshinov, V.N., 1986. *Marxism and the philosophy of language*. Harvard University Press, Cambridge, MA.

Vrikki, M., Wheatley, L., Howe, C., Hennessy, S., Mercer, N., Vrikki, M., Wheatley, L., Howe, C., Hennessy, S., Mercer, N., Wheatley, L., Howe, C., Hennessy, S., 2018. Dialogic practices in primary school classrooms Dialogic practices in primary school classrooms. *Lang. Educ.* 0, 1–19. https://doi.org/10.1080/09500782.2018.1509988

Walker, J., 1994. The body of persuasion: A theory of the enthymeme. *Coll. English* 56, 46–65. https://doi.org/10.2307/378216

Walton, D., 2008. The three bases for the enthymeme: A dialogical theory. *J. Appl. Log.* 6, 361–379. https://doi.org/10.1016/j.jal.2007.06.002

Walton, D., 2007. The speech act of clarification in a dialogue model. *Stud. Commun. Sci.* 7, 165–197.

Walton, D., 2001. Enthymemes, common knowledge, and plausible inference. *Philos. Rhetor.* 34, 93–112. https://doi.org/10.1353/par.2001.0010

Walton, D., 1992a. *The place of emotion in argument*. Pennsylvania State University Press, University Park, PA.

Walton, D., 1992b. *Plausible argument in everyday conversation*. State University of New York Press, Albany, NY.

Walton, D., 1987. *Informal fallacies*. John Benjamins, Amsterdam, Netherlands.

Walton, D., Krabbe, E., 1995. *Commitment in dialogue*. State University of New York Press, Albany, NY.

Webb, N.M., Franke, M.L., Ing, M., Wong, J., Fernandez, C.H., Shin, N., Turrou, A.C., 2014. Engaging with others' mathematical ideas: Interrelationships among student participation, teachers' instructional practices, and learning. *Int. J. Educ. Res.* 63, 79–93. https://doi.org/10.1016/j.ijer.2013.02.001

Wegerif, R., 2019. Dialogic Education, in: Noblit, G. (Ed.), *Oxford Research Encyclopedia of Education.* Oxford University Press, Oxford, UK. https://doi.org/10.1093/acrefore/9780190264093.013.396

Wegerif, R., Doney, J., Richards, A., Mansour, N., Larkin, S., Jamison, I., 2019. Exploring the ontological dimension of dialogic education through an evaluation of the impact of Internet mediated dialogue across cultural difference. *Learn. Cult. Soc. Interact.* 20, 80–89. https://doi.org/10.1016/j.lcsi.2017.10.003

Widdowson, H.G., 1979. *Explorations in applied linguistics.* Oxford University Press, Oxford, UK.

Wolfe, E., Moulder, B., Myford, C., 2001. Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *J. Appl. Meas.* 2, 256-280.

Zahavi, D., 2014. *Self and other: Exploring subjectivity, empathy, and shame.* Oxford University Press, Oxford, UK.

Zahavi, D., 2008. Simulation, projection and empathy. *Conscious. Cogn.* 17, 514–522. https://doi.org/10.1016/j.concog.2008.03.010

What a Difference Depth Makes

Dina Mendonça

Mendonça, D. (2019). What a difference depth makes.
***Revista de Filosofia Aurora*, *31*(54), 671-694.**
https://doi.org/10.7213/1980-5934.31.054.DS01

Emotional depth is an aspect of Emotion Theory that is still to be fully explored. Though depth is undeniably important for phenomenal experience and emotion (Gaebler et. al. 2013) it remains an issue to be thoroughly researched. Partly because often the criteria for identifying and understanding emotional depth is simplistically trapped in circularity by claiming that deep emotions are important because they refer to deep and important aspects of people's lives. This paper offers a novel way to capture emotional depth claiming that the narrative structure of emotions holds the key to understanding why some emotions are at the surface while others are hidden, and also the way to fully comprehend why some emotions are more decisive, deeper and profound than others.

Emotional Depth

Mind has depth. Though the fact is undeniable and depth is crucial for understanding experience (Gaebler et. al. 2013), reflection upon its place in understanding emotions, mind and experience is sparse. Accordingly, Danto points out in "Deep Interpretation," (1981)

> surface interpretation, which we are all obliged in the course of socialization to become masters of, has been extensively discussed by philosophers in the theory of action and in the analysis of other languages and other minds. But deep interpretation has been scarcely discussed at all. (Danto 1981, 695)

Similarly with emotions: though depth is an important part of understanding them and their connection to other aspects of the mind, reflection and discussion about emotional depth is rarely the focus of philosophical work. Echoes of this assessment can be found in Cataldi's book *Emotion, Depth, and Flesh: A study of Sensitive Space* (1993), when she writes, "[d]espite our implicit understanding of the phenomenon of emotional depth, there is next to no philosophical literature devoted to the topic." (Cataldi 1993, 1) She names two exceptions: Max Scheler's discussion of the stratification of emotional life in *Formalism in Ethics and Non-Formal Ethics of Values* (1973), and John Dewey's early work on the deepening of emotional feeling discussed in his *Psychology* (1887). More than twenty years later we can add two references to this list: Cataldi's book and Pugmire's *Sound Sentiments* (2005). In sum, even though understanding emotional depth is crucial to the field of emotions and is in part recognized in the literature because of its impact on mental health and education (e.g. Ben-Ze'ev 2000, 485; Baier, 1990, 4), it is still open for research how emotional depth is to be understood, fostered, and enhanced for a complete picture of emotions and of the mind.

At first sight the difference between deep and shallow emotions is given by how certain emotions are connected to crucial and deep structures of life and the self. Pugmire writes that, "depth depends at least on how much of a person's life is affected by what evokes the emotion (e.g. Fear of what threatens everything I have striven for)." (Pugmire 2005, 43) That is, the difference is grasped by acknowledging that if these deep emotions were not in place it would change drastically the meaning of life, and the identity of the self. In contrast, shallow emotions are connected to things that are less important, and which would not be missed if they ceased to exist or if they were altogether absent. This means that fear will be deep if something threatens one's life, or if there is a threat to a crucial relationship or value held by the self; and the fear will be shallow if it arises with less important issues such as feeling scared of being overly dressed for a certain occasion. That is, even though we use the same word to describe the emotional experience, it is unquestionable that in one instance the fear is deep and in the other it is shallow.

Cataldi claims that we do not experience "deep emotions over things that are manifestly superficial," (Cataldi 1993, 2) and consequently emotional depth is already given and differentiated in our life. Cataldi also thinks that the difference between deep and shallow emotions is mirrored in our vocabulary such that "cruelty is deeper than spite, awe is deeper than admiration, sorrow is deeper than sadness, joy is deeper than gladness, reverence is deeper than respect; and so forth." (Caladri 1993, 7) She would perhaps argue that the use of fear in the example above of being overly dressed for an occasion is a misuse of language, and that one could more adequately describe the emotional experience as a slight sense of anxiety about fitting in. According to Cataldi, all that is required for dealing with emotional depth, and to differentiate it from shallow emotional experience, is to trust the use of language because people know that they should use remorse instead of regret when the issue is deeper.

The fact that emotional depth can be given in language means that somehow it must also be already present in the phenomenology of emotions, and one can count on people's ability to distinguish the deep emotions from the shallow ones and to name them adequately. And, this is why people recognize when something emotionally deep has happened to them by its impact, longer duration and greater intensity.

However, sometimes people are not capable of properly naming their feelings (e.g. Pugmire 1994, 108; Haybron 2007, 394). And even thought there are some particular emotions that are not possible to experience superficially for "we cannot be rather grief-stricken, hate half-heartedly, or be a touch ecstatic," (Pugmire 2005, 31) some emotions can both be experienced superficially on some occasions and more deeply on others, as when a person "can be furious or just annoyed, and either slightly or desperately worried." (Pugmire 2005, 31) Accordingly, Pugmire states that emotional depth requires a more rigorous description than the criteria given by language refereed by Cataldi. This explains why a full grasp of emotional depth requires education, training and sometimes even professional help from experts of the mental health sector.

Pugmire proposes three conditions that make an emotion deep: (1) it must be cognitive (believed instead of entertained), (2) it must hold a certain range of response regarding the wholeness of mind, and (3) there must be a harmony "between the significance I give to something (in (1) and (2)) and its actual magnitude." (Pugmire 2005, 49) The advantage of Pugmire's criteria is that it moves the discussion beyond the scope of psychological and phenomenological descriptions of experience of emotion offered by Cataldi, and opens an understanding of emotional depth capable of a description that "depends on more than personal susceptibility and is not just a property of emotional experience on its own." (Pugmire 2005, 33) Thus, while Cataldi's understanding of depth only requires trust in the subjective phenomenology of the emotional landscape and a good grasp of emotional vocabulary, for Pugmire depth depends on the world such that "the character of that world may decide the quality of responses available to an individual." (Pugmire 2005, 33) Consequently, the conditions that enable depth of emotion are "indeed partly external to the emotion; and not only partly but heavily so," (Pugmire 2005, 33) requiring an evaluation of the phenomenology of experience and of the situation in which it arises. By including the external conditions of emotional experience, Pugmire does justice to the Janus-faced nature of emotions, and adds to the picture of emotional depth this long recognized feature of emotions that "[t]hey tell us something about the world, and they tell us something about ourselves." (De Sousa 2007, 323) So emotional depth is similarly measured by combining its internal and external aspects, and therefore "emotions could acquire depth from the person's concerns (e.g. desires, needs, valuations, other emotions) in the same way they draw on depth of belief." (Pugmire 2005, 40)

Given the need to incorporate the elements of the subject and the elements of the world, as well as a sense of harmony between them, Pugmire concludes that an emotion cannot be deep unless it is morally adequate (Pugmire 2005, 63), turning the insightful recognition of emotion's Janus-faced nature into the conclusion that depth of emotion reflects the deep elements of life. He writes,

> Depth of emotion, then, may reflect excellence of character. Through such emotion a person participates as fully as possible both in his own life and in that of the world through which he passes. By the same token, deep emotion is a reflection of the world. If the elements of someone's life or of the world are themselves shallow, so are their proper emotional resonances (Pugmire 2005, 64).

The end result of Pugmire's definition points to a similar implication Cataldi's: namely, that it is not possible to feel deeply about superficial things, nor feel superficially about deep things because "profundity of emotion depends on the quality of its subject matter." (Pugmire 2005,

58) That is, even though Pugmire's requirements for emotional depth are more rigorously demanding it is still the case that depth appears trapped in a circularity in which deep emotions are important because they refer to deep and important aspects of people's lives. And except for those who hold excellence of character, most people will oscillate between sometimes making the proper and adequate judgment with the right intensity and with perfect harmony between significance and magnitude, and other times they will not succeed in attaining such clarity of understanding on the level of depth of their emotions. Thus, the issue returns to the same difficulty: people are not always clear about what counts as depth in their emotional experience, even though there are some common sense guiding assumptions about when emotional depth is present.

Take, for example, how the temporality of emotions gives us some insight into their depth. When emotions are brief or short they can more easily be labeled as shallow, in contrast with enduring ones. Nevertheless, it remains the case that duration is not a sufficient condition to label an emotion as deep. Sometimes one may feel a deep emotion even though it does not persist for a very long time. And sometimes a superficial emotion can last because, as Pugmire points out, "[j]ust as a belief can endure as a dogma, an emotion can persist as a kind of psychological reflex, by dint of habit." (Pugmire 1994, 37) Similarly with intensity: it is not a sufficient condition for emotional depth for "[a]n access of intense feeling, such as a tantrum, the frights, or giddy elation, may just be an excess of it." (Pugmire 2005, 34) This is because the phenomenology of depth is not tied to a particular and specific experience since "experiences that have a particularly high or low phenomenal depth are not just more or less intense but qualitatively different," (Gaebler et al. 2013, 271-272) and consequently emotional depth is also qualitatively different from the experience of emotional intensity.

Acknowledging that deep emotions are more enduring than superficial ones, and that they are more meaningful, more valuable, more important and more essential are not sufficient criteria for identifying emotional depth. And it is possible to imagine a superficial person having a very deep emotional tie to their superficial experiences, as when some people express great emotional intensity when they worry about the next episode of this or that series on television, or about make up, or about the brand of clothes they are wearing. Pugmire may reply that the person who feels such emotions is not holding the appropriate harmony between the personal given significance, and the actual magnitude of the importance of these issues. However, it will be hard to show that these persons' lack of magnitude when they experience it in such a way that it acquires such magnitude. Not being able to clearly explain to others why and when emotional depth is misplaced means emotional depth's intricate nature is not fully grasped to be communicated, making it difficult to show why and when people should revise their emotional shallowness and foster deeper emotions.

Invoking Nagel's essay on "What is it like to be a bat?" (1974) Danto insightfully points out this limitation about naming depth based on its phenomenology. He explains that though we cannot know, as bats know, what is it like to be a bat, it is also the case that, "bats, if they have depths, are no better situated than we for knowing what it deeply is to be a bat." (Danto 1981, 694) This means that when we are in the face of the deep we may not clearly see it as deep for "[i]n the depths there is nothing that counts as being there." (Danto 1981, 695) Perhaps we cannot completely grasp emotional depth's full conditions because experience of an emotion does not come with a clear and distinct label of its depth, and consequently, even Pugmire's more rigorous description of emotional depth cannot acquire sufficient precision.

Nevertheless, experience also grants some recognition of the difference and, at least in education, people guide children's emotional development (Kristjánsson 2002, 18; 2013, 192), often by pointing out to children that some things are not worth crying about in comparison with others. The parent who pays less attention to a fit of anger about something superficial, and gives attention and care upon anger about a felt injustice is guiding the child to distinguishing the superficial from the deep. Writing about emotional education and emotional development, De Sousa suggests that emotional learning is similar to aesthetic education (De Sousa 1990, 436). Building upon this suggestion it is possible to propose that just as aesthetic sensibility can be fostered by experiencing art in a variety of settings and formats, which offers ways to compare and contrast aesthetic qualities, emotional education requires incorporate the identification of deep emotions in contrast with what is superficial by placing emotions into perspective and thus adding meaning to their distinction. This means that talking and reflecting with others about how emotions arise in stories aids grasping and understanding them (Lipman 1995, 5; Kristjánsson 2001, 10).

Danto describes a similar insight about how contrast between surface interpretation and deep interpretation provides an added perspective upon action. He shows how when we adopt the theoretical posture of taking up a practice of archeology about actions, we get a more complex picture of action in which two levels of interpretation are given. As Danto explains, we have a picture in which a certain action is done and described in a certain way (a), and when we adopt a certain new theoretical posture a new different description of it appears (b). The meaning of the action is only brought about when we understand that in doing action (a), action (b) is what really is being done, such that "it is hidden from the a-doer that he is a b-doer. A deep interpretation of a identifies it as b, whereas a surface interpretation identifies it as a." (Danto 1981, 698) This means that the deep interpretation gives another meaning to the superficial one and yet one cannot see it without having the superficial one as well because one cannot insightfully claim that the a-doer is really a b-doer without the two interpretations. If we transfer Danto's explanation about interpretation to emotions, we can say that once we have access to deep emotions that underlie an experience, the superficial ones change their meaning and become less crucial and important by comparison, even if they persist in intensity and continue to appear in experience. However, without the superficial emotional experience there is no way to grasp what are the deep ones and how they are revealing, and it is impossible to understand their meaning and the renewed meaning of the superficial ones.

To clarify this point let me illustrate it with a metaphor about the water at the bottom of a well. Imagine someone looks into a well from the top. The person may not be able to see what there is at the bottom. The person can perhaps verify the existence of water by throwing a stone in the well, and finding out through the sound of the splash if it is worth the effort of lowering a bucket. Once the bucket has been lowered and water from the bottom raised to the top, the water will look no different from the water one obtains from the surface of a lake, and at most the person will recognize that there are two possible levels at which one can find water, and that the bottom of the well requires a more complex process to be reached, in which a person can learn more about their abilities and about the dimension of the world. Discovering our deep emotions is similar: the process of identifying emotional depth can reveal aspects of the world and abilities to interact with it. It is recommended because it enables the ability to distinguish which emotions to take seriously and which ones to treat with lightness by putting emotions in perspective by establishing a contrast.

One of the ways in which people learn to build emotional perspective – of placing deep emotions in contrast with superficial ones – is by listening and being asked to describe things in narratives because when people describe a situation someone else or themselves are living, they construct a narrative that places emotions in perspective. For example, a story about how a person can laugh at a silly joke among friends while still feeling intense sadness and the beginning of grief for the cousin who died in a motorcycle accident will identify that the laugh is not a deep sense of joy but instead a way to show the appreciation for friendship in a deep, difficult confrontation with the loss of a family member and with death. That is, stories put emotions into perspective by indicating that not all emotions are visible all the time, and that some emotions are at the surface while others are hidden for a variety of reasons. This enables understanding how some emotions are more crucial than others, and that some are more decisive and deeper than others in what concerns the self. In addition, depth comes with the recognition that there are different levels of depth and superficiality and emotional depth is a matter of degree of depth and not a quality that an emotion either has or lacks. Consequently, stories also reveal different dimensions of depth and how these are related to each other.

Narrative Structure of Emotion

The ability of stories to put emotions into perspective is recognized by different philosophers who work on Emotion Theory when they claim that emotions have a narrative structure. Here are some examples: Martha Nussbaum writes, "[e]motions, we now can see, have a narrative structure," (Nussbaum 2001, 236) pointing out how the cognitive emphases in philosophy of emotions has shown the crucial importance of emotions' narrative structure. Nussbaum thinks that without this narrative history we cannot have a complete understanding of emotions because, she writes, "[t]he understanding of any single emotion is incomplete unless its narrative history is grasped and studied for the light it sheds on the present response." (Nussbaum 2001, 236) And she reinforces the importance of the narrative stating that certain things about emotions can only be grasped by it. She writes "[t]his is what Proust meant when he claimed that certain truths about the human emotions can be best conveyed, in verbal and textual form only by a narrative work of art." (Nussbaum 2001, 236) In *Upheavals of Thought: The Intelligence of Emotions* (2001), Nussbaum suggests that the narrative mirrors the structure of an emotion and that only an artistic version of the sequence of the relevant events, actions, thoughts, and feelings can truly grasp emotion's structure arguing that the narrative structure of emotion is complex and refined.

Annette Baier also refers to the history underlying an emotion when she states that, "the full content of an emotion always refers to the past, whatever else it refers to. Emotions are history-laden states of mind." (Baier 1990, 18) She adds another criterion to emotions' narrative structure when she further explains that "[c]hildhood, adolescence, youth, maturity, old age, will be *relevant* to the appropriateness of given emotions, in a way it is not to either the truth or the reasonability of holding particular beliefs." (Baier 1990, 19) That is, Baier thinks that narrative structure of an emotion must also take into consideration the temporal path of the person who feels the emotion and somehow provide the temporal perspective of the person who experiences the emotion.

De Sousa also points out emotions' narrative structure when, in *The Rationality of Emotions* (1987) he argues that, "we acquire the capacity to talk about emotions in terms of the stories that give rise to them." (De Sousa 1987, 183) De Sousa thinks that while learning these stories we are learning paradigm scenarios that are then associated with our vocabulary of emotions. The backbone the notion of paradigm scenario is the story that is somehow present in paradigm scenarios such that "by the time toddlers are four to five years old, they have a very good sense of what kinds of stories lead to what simple emotions." (De

Sousa 1987, 183) The use of the term 'narrative' in De Sousa's work is closer to a simple notion of a sequence of events just as presented in a children's story and it does not require the artistic, creative touch that Nussbaum seems to demand. However, the notion of paradigm scenarios aims to be a technical term and consequently these paradigm scenarios do not need to be interpreted but are used to interpret situations that happen in life and can be turned into narratives (De Sousa 1990, 438). It is important to note that De Sousa stresses the open ended character to emotions' narrative structure pointing out that "a paradigm can always be challenged in the light of a wider range of considerations than are available when the case if viewed in isolation," (De Sousa 1987, 187) giving stories and experience the capacity to increase complexity of already existing paradigm scenarios.

Although the previous examples are a good sample of many different thinkers who consider emotion to have a narrative structure, it is also clear that they use the notion of narrative in different ways. It is beyond the scope of this paper to verify if these different ways are contradictory or if, on the contrary, they can all be accommodated. Nevertheless, it is common to all authors that the notion of a narrative structure of an emotion means that there is a sequence (events, actions, thoughts, feelings), and that the sequence takes the format of a story that provides a temporal organization that gives insight about the nature of an emotion.

The narrative structure of emotions is also a way to better understand emotions' connection to the self. The insightful power of narrative for self-identity is testified by how narratives are one of the best ways to understand us and other people, and to make sense of our own actions and of others (Gallagher 2006, 228). Peter Goldie, who also speaks of the narrative structure of emotion, explains precisely how this narrative structure of emotion enables an important connection to the narrative of life. He claims that it is the underlying narrative structure of emotion that enables an understanding of the complexity of an emotional state because narrative involves a variety of elements described in an organized whole, in which connections between different parts and different elements among themselves and to the whole are made visible. He further writes that, "[a] particular emotion has a complex narrative structure which unfolds over time," (Goldie 2002, 101) and explains that an emotion has several emotional episodes which include a variety of different elements such as perceptions, thoughts, bodily changes and feelings. That is, emotion has a narrative structure because life has a narrative structure. Goldie writes, "[a] person's emotion will comprise elements or episodes which are bound together as part of a *narrative structure* which makes best sense of this aspect of the person's life." (Goldie 1999, 395)

Pointing out that emotions are best understood in light of a narrative structure does not mean this is the only way to understand them. As Shaun Gallagher points out, the insightful nature of narrative and the recognition that emotions are well suited for being captured by a narrative format "doesn't mean that our understanding of others requires an occurring or explicit narrative story telling: but it does require the ability to see/to frame the other person in a detailed pragmatic or social context, to understand that context in a narrative way." (Gallagher 2006, 226) The importance of how narrative best grasps the dynamic nature of emotion is revealed also by how they provide perspective upon emotions showing more clearly the contrast of the deep from the more shallow emotions.

The suggestion that narrative is a tool for seeing depth may inadvertently imply that although depth is not installed by narrative, it is captured by it. However, several problems and objections can be raised. The first objection is that it is easily seen how in the animal kingdom, at least for mammals, some things are more crucial and important than others, and therefore some things will matter emotionally more deeply than others. For example, nonhuman animals grief (Nussbaum 2017, 138) and the testimony as well as the experience of animal researchers "supported by empirical data, show that many animals experience deep emotions ranging from joyful glee when playing to bereavement, grief, and depression over the loss of a mate, child, or other friend." (Bekoff 2002, 103) Yet, animals do not seem to have access to narratives the way humans do, and this means the way narratives captures perspective is an incomplete description of how to understand emotional depth.

A second objection may be raised arguing that it is when words fall short of being able to express the emotional experience that something deep has been felt. This happens precisely because "the feelings evoked are so very deep (perhaps primal) that there are no words rich enough to convey what we feel." (Bekoff 2006, 280) The suggestion that narrative captures emotional depth risks not only being incomplete, but completely off target for it is what is not possible to capture by words or description what signals emotional depth.[1]

Finally, a last general objection may point out that the notion of narrative itself is a problematic concept with many difficulties and troubles that plague it as a theoretical tool (Currie 1998, 654). One difficulty are the rival definitions of narrative in the literature, and the previous description of how philosophers of emotion differently describe the narrative structure of emotion is a natural consequence of this plurality of definitions. For example, Levingston describes three different definitions of narrative (Levingston 2009, 26): the first definition is the minimal conception of narrative, which entails the presentation of one event minimum or one change of affairs, and can easily be recognized in the previous description of the narrative structure of emotion as the description offered by De Sousa. The second one, which Levingston describes as a slightly more complex notion of narrative which incorporates a series of casually related events, and finally, the third and final notion of narrative, an even more complex notion which entails a discourse from the perspective of an agent in which we can identify goals, obstacles and a sequence of a path to achieve the goal and where the narrative explains the causal structure of the sequence of events (Levingston 2009, 26). Levingston criticizes all three definitions writing that the first is too vague, and easily raises disagreement about what counts as a narrative and what are its minimal criteria. The second definition requires an explanation and justification and a good theory of metaphysics of causation, which is often not offered by these same authors (Levingston 2009, 27). The third definition has been criticized by making the problem solving description the central trait of narrative (Bruner 1991, 10). In face of these problems, Levingston concludes that,

[1] Many thanks to Danil Razeev, Daria Chirva and Maria Sekatskaya for raising these problems at the presentation of an earlier version of paper at the international Conference "Ontology of Subjectivity: Selves, Persons and Organisms " at the Institute for Philosophy of St. Petersburgh State University, (September 2015).

despite its continuous use, narrative remains a deeply ambiguous term (Levingston 2009, 28).

In addition to the problem of definition, there are other difficulties to consider when we take emotion to have a narrative structure. The suggestion of a narrative structure of emotion raises the question of who should be the narrator of a specific emotion. That is, it is not clear whose perspective is taken when telling the narrative: if it is the individual who experiences the emotion, or if it is an external perspective of someone who sees another experiencing the emotion, or if it is some type of god-like position from which the narrative structure of an emotion is described, or if it is an even more abstract notion of narrative in which we consider this narrative as a sort of story that occurred and that no one is telling it (Currie 1998, 655). This issue raises a subsequent difficulty of wondering who the narrative is addressed to and to question whom is the "implicit 'narratee' of figure to whom the narrative is addressed." (Levingston 2009, 27) A third difficulty is given with the issues regarding the format of such narratives. It is not clear if it is a common format for all emotions or if different emotions require different narrative schemes. Clearly, narratives are commonly linear and take things one after another. That is, when we consider the narrative structure of emotion we may be promoting a linear temporal description such as, for example, first the subject sees the snake, then feels fear, then the subject runs or, to use James' modified version of sequence, first the subject feels bodily modification in face of an event, then perceives the modifications, then feels fear and then runs (James 1984, 189-190). However, one can argue that in the case of emotions one needs a much more complex format, closer to Nussbaum's description, in which temporal jumps are allowed and cherished for accurate descriptions of different emotions and subtle details of phenomenological experience and interpretation of the experience. In sum, what type of format should be adopted to understand emotions' narrative character is not clear. Even if one would adopt a position in which a multiple of narrative formats would be allowed, an explanation of the selection and variation of narrative format accepted would still be required.

Profundity

The reply to the above objections and problems lies in understanding that there is a difference between deep and profound. So far we have acknowledged that deep emotions are important because they grant us a sense of emotional perspective that is absent from the mere horizontal description of emotions, and that this added dimension of thickness of emotions grants perspective upon the world and of ourselves. However, the contrast in perspective identified reveals something else about emotional depth: what is initially hidden is not necessarily equivalent to profound. This is why Danto writes, "[d]epth, needless to say, has little to do with profundity." (Danto 1981, 695) One can have deep emotions that are self-revelatory but that one would not necessarily describe as profound.

Cataldi and Pugmire both use deep and profound as synonyms because they take deep emotions as being unquestionably connected to things worthy of being profound. The reflection undergone earlier shows that there is a distinction between the two because one can only recognize an emotion as deep if it is in contrast with what is superficial. When the contrast is established, what is hidden becomes more visible, and it is also possible to see that sometimes what is hidden is not important. Thus, what is deep and hidden is not immediately equivalent to what is profound. However, revealing what is hidden and putting emotions in perspective with the contrast between superficial and deep is crucial to help unmask what is important, central and worthy of being profound. This means that uncovering the deep can be revelatory to the point of transformation. Once deep emotions are revealed, people are in a position to acknowledge what they think is crucial and important. The suggestion of gaining perspective by contrasting superficial and deep shows that the deep is transformative, not because it reveals what is already profound, but because it can be an important step towards attaining profundity.

Take for example an adult who is afraid of water because of a traumatic experience in the early years of their life. Knowing that the fear was acquired at a younger age that one has no memory of may explain the sense of depth of its presence, but it would not be taken as a profound insight about water, about fear or about the relationship between persons and water. Now imagine that this person does not know about this traumatic experience and has no recollection of the event. The deep sense of fear of water is present and acts invisibly, appearing as a strange sense of anxiety around water. If the person discovers the occurrence of a traumatic experience with water in their first years of life, the deep emotional reaction may become more visible and its force no longer works in the dark. The person who discovers that their deep fear of water is due to a traumatic experience in their childhood may not be able to annihilate feeling scared near water but the person will know that such intensity is not profound for it is not linked to the very nature of water, the very nature of personhood, nor to a crucial connection between persons and water.

The deep feeling of fear will be recognized as somewhat accidental even if it does not feel that way, and its felt appearance of importance has to do with the specific genealogy of the feeling, and should consequently be treated with some lightness concerning the evaluation of danger in face of water. Nevertheless, the felt importance also reveals a profound commitment to the importance of life and this may bring a whole different type of focus to the person who suffers from this fear. Discovering the source of the fear by the narrative about one's childhood, and the recognition of how intense is the commitment to life can be transformative. However, narratives of the past are not always available and sometimes the ones that are available do not grasp what is hidden as to allow the distinction between what is deep and what is profound. In fact, the source may not be a clear-cut story as is suggested in the given example, and the source may be much more subtle and complex and not grasped by the identification of a single reason. Nevertheless, it is by exploring the power of narratives that the distinction between deep and profound is made clear, and it also enables identifying what should be dealt lightly and what should be dealt with commitment. How a story about a person and their feelings and emotions is told helps that person and others around that person to capture what is fundamental, and distinguishes it from what is superficial.

There are a variety of reasons why narratives have this power to unlock insight. The first reason is that the notion of narrative is very suited for emotions (Gallagher and Hutto 2006) because it is particularly suited to the description of agency (Currie & Jureidini 2004, 415). The way narratives are selective and perspectival (Lamarque 2004, 398) allows us to integrate the selective and perspectival character of a felt emotion and also make it reasonable how emotional distortion can be evaluated differently for "[a] novelist who deliberately distorts historical events for artistic ends is not subject to the same judgment as the historian whose distortions are due to ignorance, bias or deception." (Lamarque 2004, 399) Similarly, distortion of emotions is sometimes a crucial element to fully understand them.

The second reason is that narrative demands attention to detail, and in emotional experience details matter greatly. A pang of jealousy felt by a spouse can be caused by a small detail in a conversation, as when a husband looks subtly away when the name of a person is implicitly present in a description of an event at work. The notion of narrative enables us to capture this sensibility to details that is part of emotional life. For example, jealousy can be triggered by the way the story is being told, and the way details rise to the surface and become crucial. Fiction writers are masters of naming out details and, even more to the point, of disguising their relevance for the reader such that their crucial meaning is later found out in the story - detective stories are the easiest and clearest examples.

The third reason why narratives foster insight about depth is that they seem to be capable of capturing emotional experience in its movement such that there is nothing outside the story for an emotion because a "story begins with the circumstances that initiated some affect, or sequence of affects, and it ends when the emotional sequence is in some way brought to a close" (Velleman 2003, 14). As Jerome Bruner writes, narratives are a way to capture 'lived time' (Bruner 2004, 692) and in emotion that is really crucial for it is not just the felt emotion that captures the emotional episode but the antecedent conditions and the subsequent actions, emotions and events that make justice to the full understanding of jealousy or of an instant of fear. This ability of narrative to capture lived time allows us to recognize the relevance of the type of self in the emotional episode, how different contextual conditions have decisive impact on felt emotions, and how other emotions interact in the structure of a specific emotion.

The fourth reason why narrative unlocks insight for emotional experience is that narratives provide a mode to compare, contrast and connect the emotions of others and our own. In this regard Goldie writes,

> “[i]n understanding through reason and imagination, we use the information which we have about another person to, as it were, piece together or fill in the gaps in the narrative structure by bringing to light the episodes of the emotion in a way which will make best sense of this part of his life; ‘piecing together’ and ‘filling in’ the narrative structure are, in fact, just the appropriate metaphors.” (Goldie 1999, 397)

Goldie distinguishes several different ways in which people grasp other people’s emotions (Goldie 1999, 395). First, one can understand and explain another’s emotions by understanding the content of a story (e.g. when someone tells a story of seeing a snake while camping the listener can identify that it is a scary experience). Second, one can be emotionally caught by someone else’s emotion, as when someone begins to feel scared by seeing someone in a campsite running in a scared way to the river and thus becomes alert to possible signals that trigger fear in that context. Third, one can centrally imagine the other’s emotional experience from the internal perspective where the person tries to imagine being the other person who is experiencing the emotion. Fourth, one can imaginatively put oneself in another’s shoes and try to imagine what one would do if placed in the situation of another and narrating the same event as it happened to another by being oneself as the subject of the emotional experience. The difference with previous description is that here one brings a deliberate mixture of one’s own character into the process. Finally, one can focus on the outcome of the event and recognize the other person’s difficulties, and while experiencing feelings of distress for the other person become motivated to alleviate those difficulties in some way.

All of these ways to grasp someone else’s emotions show that without the tool of narrative it would be much harder to explore different modes of thinking about other people’s emotions. Ultimately when we recognize how narrative provides a tool for such differentiation we better understand how narratives have a powerful effect on the imagination and, in doing so, can also affect motivation (Currie & Jureidini 2004, 419) and the way we think about our own emotions and recognize that a narrative provide not the "familiar patters of *how things happen*, but rather to familiar patterns of *how things feel.*" (Velleman 2003, 19)

Finally, the last reason that makes narrative insightful is that it is a way to handle emotional complexity and emotion research needs to "take complexity seriously rather than ironically and acknowledge it by default." (Colombetti 2005, 123) Emotional complexity is given by narrative because it captures emotion's dynamic nature, and the interaction of the person who feels it with the world. Perhaps more importantly, the story telling cannot be substituted by a sequence of reasons giving that come out of the narrative. The details, tones and nuances offered by the stories go beyond the list of reasons they offer[2]. This is also why they cannot be tailored to the persons' needs even though they are an intrinsic part of they are. As Whollheim writes about it,

> This interaction is embedded in the narratives that we associate to our emotions, and in these narratives, conscious or unconscious, lie the identities. But we must not think that these narratives are stories that we can make up at whim or at will. They are probably as deep as anything that we know about ourselves (Wollheim 1999, 224).

In sum, to benefit from gaining perspective upon emotions we cannot simply invent narratives. When stories are not available, or when the ones available do not enable a way to distinguish the deep from the profound, it is necessary to enter the process of story telling - to explore the various dimensions of narrative that can provide new possibilities and investigate different takes - to see what is profound. Though none of the philosophers of emotions who identify the narrative structure of

[2] Many thanks to Prof. Douglas Cairns and his question on the presentation of the paper at the "The Art of Feeling. Emotions Across Disciplines and Genres" in Lisbon (June 2019), and subsequent lively discussion for refinement of the argument.

emotion consider it, we may easily imagine that there can be two different tokens of the same narrative just like we have "two photographs of the same scene from different angles." (Currie 1998, 656) This means that the same narrative structure of emotions can be differently described with variation in styles, perspectives and focuses of attention.

The process can be long, and it also has no guarantee that exploring story telling will for sure end up making people able to separate the wheat from the shaft. However, if the story telling research is complete and succeeds in uncovering the deep emotions, such that it harvests insight to distinguish it from profound, it can offer a transformative experience. This transformational move is perhaps the reason why deep is taken to be revelatory and self-revelatory, and provides a way to get novel perspectives on experience and life. Cataldi writes, "after a "deep emotional experience" we may say that we are "not the same person" or we may realize that we are beginning to see things in a different way or in a different light." (Cataldi 1993, 1)

It is the problems and difficulties of narrative that also hold its richness and can be used to distinguish the deep from the superficial. And it is this added step that distinguishes the effect of emotional depth in animals. Animals may have sufficient emotional complexity to have superficial in contrast with deeper emotions, but without the use of language that narratives require, they do not appear to have the voluntary transformative experience of uncovering deep emotions to identify and commit to the profound ones. Animals do not need to reach the peaceful sense of the profound because they do not appear to have to choose among values, or establish a hierarchy of values. Life is the ultimate value. For humans the notion of a good life requires that values be discussed, compared and chosen. More importantly, humans do not do this process in isolation, and the search for what is a 'good life' is a social and cultural ongoing conversation in which narratives play a decisive role in sharing, communicating and interpreting experience.

People reach the stage of sharing deep emotions without speaking not because deep emotions do not require words but because they have reached the sense of deep emotions and are able to share it without having to tell a story. That is, as a species we count on other people's ability for storytelling to discern the deep from superficial emotions. The profound needs an added step and can perhaps more easily be identified in a narrative format when someone has to retell a story such that the contrast between the deep and the superficial mirrors a choice of values.

The reflection of emotional depth opens a totally new field of research for emotion theory. Researches can be sure of its reality for it is because there is superficial versus deep emotions that emotional depth is revealed as such. And although uncovering the deep is difficult and requires work and progress to show what is hidden it can be hard, slow and challenging. However, once the deep is uncovered it is possible to separate it from the profound and then open the possibility for transformative experiences. When the deep is identical to profound no transformation necessarily happens for the only thing that occurs is the addition of information and insight about knowing oneself and the world a little better confirming what was already present. This means that even though diving deep into our emotions can be a way to better understand ourselves and answer the ongoing questions about our self-identity, the decisive moment of uncovering the deep is located in its distinction from the profound, and the commitment to action guided by such transformation. When profound emotions are recognized people are at peace, and while the process of uncovering the deep is turbulent and always in relation with what is at the surface, the profound is a commitment that can guide action in a calm and steady fashion. This meaning of emotional depth and its contrast with the superficial is only fully understood when a sense of the profound is uncovered and established by narrative. Ultimately this means that the ability to reach the profound is not a given, and it further reinforces the importance of acquiring and cultivating narrative literacy (Hutto 2007, 47; Gallagher and Hutto 2008, 28-35).

The ability of transformation in the identification of emotional depth and the distinction from what stands as emotionally profound has been so far absence from research on emotion and may provide a further insight into the ways in which emotions are part of rationality (Williams 1981, 29) such that we can more fully understand when emotional experience tells us something decisive about the world from when they distort our view of things (Goldie 2004, 249).

Acknowledgments

This research work is supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., in the context of the celebration of the program contract foreseen in the numbers 4, 5 and 6 of article 23.º of D.L. no. 57/2016 of 29 August, as amended by Law no. 57/2017 of 19 July. Many thanks for the insightful informal group discussion at EPSSE Workshop "The Meaning of Moods" at Landgut Castelen in Basel (Dec. 2015) with Achim Stephan; Valentina Petrolinic; Ruth Rebecca Tietjen; Eva Weber-Guskar; Daniel O'Shiel. I am grateful to Danil Razeev, Daria Chirva and Maria Sekatskaya for raising problems at the presentation to an earlier version of paper at the international Conference "Ontology of Subjectivity: Selves, Persons and Organisms" at the Institute for Philosophy of St. Petersburgh State University, on September 2015. Also thanks to the audiences of earlier presentations of the paper at the 3rd Annual Conference of the European Philosophical Society for the Study of Emotions in Athens (June 2016), and at the "The Art of Feeling. Emotions Across Disciplines and Genres" in Lisbon (June 2019).

References

Baier, A. "What Emotions Are About" *Philosophical Perspectives,* 4 (1990): 1-29.

Bekoff, M. *Animal passions and beastly virtues: Reflections on redecorating nature.* Philadelphia: Temple University Press, 2006.

———. *Minding Animals: Awareness, Emotions, and Heart.* New York: Oxford University Press, 2002.

Ben-Ze'ev, A. *The subtlety of emotions.* MIT Press: Cambridge, Ma., 2000.

Bruner, J. "Life as Narrative" *Social Research* 71 (2004): 691-711. Originally published *Social Research* 54 (1987): 1-17

———. "The Narrative Construction of Reality" *Critical Inquiry* 18 (1991): 1-21

Cataldi, S. L. *Emotion, Depth, and Flesh: A Study of Sensitive Space Reflections on Merleau-Ponty's Philosophy of Embodiment.* Albany: State University Press, 1993.

Colombetti, G."Apraising Valence" *Journal of Consciousness Studies*, 12 (2005): 103-126.

Currie, G. "Narrative." In Edward Craig (ed.) *Routledge Encyclopedia of Philosophy*, vol.6, London & NewYork: Routledge, 1998. 654-657.

Currie, G. and J. Jureidini. "Narrative and Coherence" *Mind & Language*, vol. 19 (2004): 409-427.

Danto, A. C. "Deep Interpretation" *The Journal of Philosophy*, Seventy-Eighth Annual Meeting of the American Philosophical Association Eastern Division, 78 (1981): 691-706

De Sousa, R. *The Rationality of Emotion*, Cambridge, Mass.: MIT Press, 1987.

———. "Emotions, Education and Time" *Metaphilosophy,* 21 (1990): 434-446.

———. "Truth, Authenticity, and Rationality" *Dialectica,* 61 (2007): 323-345.

Gaebler, M.; Lamke, J-P.; Daniels, J.K. and Walter. H. "Phenomenal Depth. A Common Phenomenological Dimension in Depression and Depersonalization" *Journal of Consciousness Studies,* 20, No. 7-8, (2013): 269-91.

Gallagher, S. The narrative alternative to theory of mind. In R. Menary (ed.), *Radical Enactivism: Intentionality, Phenomenology, and Narrative.* Amsterdam: John Benjamins, (2006): 223-229.

Gallagher, S. and D. Hutto. "Understanding others through primary interaction and narrative practice" *The Shared Mind: Perspectives on Intersubjectivity Eds.* In J. Zlatev, T. P. Racine, C. Sinha & E. Itkonen Amsterdam: John Benjamins. 2008, 17-38.

Goldie, P. "How We Think of Other's Emotions" *Mind & Language*, 14 (1999): 394-423.

———. "Emotion, Reason and Virtue" in D. Evans / P. Cruse (eds.) *Emotion, Evolution, and Rationality*, Oxford: Oxford University Press, 2004, 249-67.

Goldie, Peter. “Emotion, Personality and Simulation” *Understanding Emotions. Mind and Morals*, ed. by Peter Goldie, Aldershot/Burlington USA/ Singapore/Sidney: Ashgate, 2002, 97-110.

Haybron, D. M. “How do We Know How Happy We Are? On Some Limits of Affective Instrospection and Recall” *Noûs* 41 (2007): 394-428.

Hutto, D. “The narrative practice hypothesis” *Narrative and Understanding Persons*, ed. D. Hutto, Royal Institute of Philosophy Supplement. Cambridge: Cambridge University Press, 2007, 43-68.

James, W. “What is an emotion?” *Mind* 19 (1884): 188-204

Kristjánsson, K. “The Didactics of Emotion Education” *Analytic Teaching*, 21 (2001): 5-15.

———. *Justifying Emotions: Pride and Jealousy*. London: Routledge, 2002

———. *Virtues and Vices in Positive Psychology: A Philosophical Critique*. Cambridge: Cambridge University Press, 2013.

Lamarque, P. “On Not Expecting Too Much from Narrative” *Mind & Language*, 19 (2004): 393-408.

Lipman, M. “Using Philosophy to Educate Emotions” *Analytic Teaching,* 15, (1995): 3-10.

Levingston, P. “Narrative and Knowledge” *Journal of Aesthetics and Art Criticism* 67 (2009): 25-36.

Nagel, T. “What Is It Like to Be a Bat?” *Philosophical Review* 83 (1974): 435–50.

Nussbaum, M. *Upheavals of Thought: The Intelligence of Emotions*. Cambridge: Cambridge University Press, 2001.

———. *Political Emotions. Why Love Matters for Justice*. Cambridge, Massachusetts/London, England: The Belknap Press of Harvard University Press, 2017.

Pugmire, D. *Sound Sentiments. Integrity in the emotions*. Oxford: Oxford University Press, 2005.

Pugmire, D. “Real Emotion” *Philosophy and Phenomenological Research*, Vol. LIV, No. 1 (1994): 105-122

Scheler, M. *Formalism in Ethics and Non-Formal Ethics of Values* (trans. M. S. Frings and R. L. Funk). Evanston III: Northwestern University Press, 1973.

Velleman, J.D. “Narrative Explanation” *The Philosophical Review*, Vol. 112, No. 1 (January 2003): 1-25.

Wollheim, R. *On the Emotions* New Have and London: Yale University Press, 1999.

Williams, B. “Moral Luck,” *Moral Luck*. Cambridge: Cambridge University Press, 1981, 20-40.

Unamuno's Religious Faith in *San Manuel Bueno, Mártir*

Alberto Oya

I. Introduction

In his major philosophical work, *Del sentimiento trágico de la vida en los hombres y los pueblos* [*The Tragic Sense of Life in Men and Nations*],[1] published in 1913, the Spanish philosopher Miguel de Unamuno (1864–1936) argued for a natural, non-evidential foundation for religious faith; that is, according to Unamuno, religious faith is not legitimated because God does in fact exist, but because it is something we are naturally led to. This is why Unamuno's religious faith has nothing to do with believing, with accepting as a truth the factual claim that God exists or that the world is such and such and not otherwise. Religious faith, according to Unamuno, consists in a religious understanding of the world, in seeing the world as a sort of personal conscious being and in feeling, through the practice of charity, as if we were in a personal relationship with it — from conscience to conscience, so to say.

In 1930, seventeen years after the publication of *Del sentimiento trágico de la vida en los hombres y los pueblos*, Unamuno wrote one of his most well-known novels, *San Manuel Bueno, mártir* [*Saint Manuel Bueno, Martyr*].[2] The novel is about the fictional character Manuel Bueno, a catholic priest from a small Spanish village who, despite being unable to believe the Christian claim that there is an after earthly death life, devotes himself to the spiritual care of his people and is sanctified after his death. The aim of this paper is to show that the guideline of *San*

[1] The edition cited throughout is Miguel de Unamuno*, The Tragic Sense of Life in Men and Nations*, in *The Selected Works of Miguel de Unamuno (vol. 4)*, ed. and trans. Anthony Kerrigan (Princeton: Princeton University Press, 1972), pp. 3–358. In square brackets I cite the original Spanish text, published in Miguel de Unamuno, *Del sentimiento trágico de la vida en los hombres y en los pueblos*, in *Miguel de Unamuno: obras completas (vol. 7: "Meditaciones y ensayos espirituales")*, ed. Manuel García Blanco (Madrid: Escelicer, 1966 [1913]), pp. 109–302.

[2] The edition cited throughout is Miguel de Unamuno, *Saint Manuel Bueno, Martyr*, in *The Selected Works of Miguel de Unamuno (vol. 7)*, ed. and trans. Anthony Kerrigan (Princeton: Princeton University Press, 1976), pp. 135–180. In square brackets I cite the original Spanish text, published in Miguel de Unamuno, *San Manuel Bueno, mártir*, in *Miguel de Unamuno: obras completas (vol. 2: "Novelas")*, ed. Manuel García Blanco (Madrid: Escelicer, 1967), pp. 1127–1154.

Manuel Bueno, mártir is the expression, in fictional, non-philosophical language, of the conception of religious faith Unamuno had already defended in his *Del sentimiento trágico de la vida en los hombres y los pueblos*. By abandoning the use of philosophical jargon and expressing his view in a concrete form through the life and works of the fictional character Manuel Bueno, Unamuno is likely trying to make his conception of religious faith comprehensible to a wider audience. That this was Unamuno's intention in writing this novel is clear from what he says in the prologue to *San Manuel Bueno, mártir y tres historias más* that it should be considered as: "[...] one of the most characteristic novels of all my fictional production. And he that says fictional production —I add— also says philosophical and theological production. [...] I am aware of having put into this novel all my tragic feeling of daily life".[3]

II. God and Our Natural Appetite for an Endless Existence

Unamuno's defense of religious faith starts with the claim that all singular things naturally and primarily seek an endless existence —*i.e.*, that they all suffer from what Unamuno named as "*hambre de inmortalidad*" ("hunger for immortality"). An important point must be made here. Unamuno's reasoning does not rely on the psychological claim that we all, as an empirical fact, have the desire for an endless existence. What Unamuno's argument requires is the stronger, metaphysical claim that the most basic natural inclination (or appetite, if we are to use Spinoza's jargon) of all singular things (not only sentient beings such as us) is to seek an endless existence.[4]

[3] Miguel de Unamuno, "Prólogo a San Manuel Bueno, mártir y tres historias más", in *Miguel de Unamuno: obras completas (vol. 2: Novelas")*, ed. Manuel García Blanco (Madrid: Escelicer, 1967 [1933]), p. 1115. My translation, the Spanish text reads: "[...] una de las más características de mi producción toda novelesca. Y quien dice novelesca –agrego yo– dice filosófica y teológica. [...] Tengo la conciencia de haber puesto en ella todo mi sentimiento trágico de la vida cotidiana".

[4] Unamuno's "*hambre de inmortalidad*" has been commonly misread as referring to the psychological, empirically contingent claim that we, human beings, have the desire for an endless existence. However, Unamuno's explicit

Unfortunately, all the evidence we have goes against the claim that we will enjoy of an endless existence. As far as we know, people die sooner or later. In light of this, and by a simple induction, the only conclusion we can reasonably infer is that we too are going to die and in so doing our existence will come to an end.

There is, it is true, a long philosophical tradition which aims to prove the immortality of human beings through the use of philosophical and theological reasoning. But, according to Unamuno, these arguments are completely off the point. Even if these arguments were successful —which Unamuno argues they are not— they would only demonstrate the survival of the human soul. But we are not (at least, not only) souls: we are, as Unamuno so vividly put it, "*hombres de carne y hueso*" ("men of flesh and bone"). Therefore, these kinds of arguments cannot provide

endorsement of Spinoza's argument for the *conatus* at the very beginning of his *Del sentimiento trágico de la vida en los hombres y en los pueblos* makes it evident that he is not treating this "*hambre de inmortalidad*" as referring to a desire for an endless existence that we, human beings, have. This "*hambre de inmortalidad*" is rather a sort of primary natural tendency (*i.e.*, an *appetite* in Spinoza's jargon) to seek an endless existence, which *all* singular things essentially have. And by singular things Unamuno is referring to human beings as well as other conscious animals and *prima facie* non-sentient beings such as plants and rocks. It is interesting to note that the common failure to realize that Unamuno's reasoning does not depend on the psychological, contingent claim that we all desire an endless existence but on the metaphysical claim that all singular things seek, as their most basic natural inclination, an endless existence, is what has impeded Unamuno scholars to realize of the core and genuine aspect of Unamuno's reasoning, which is that Unamuno's religious faith is founded in our own natural condition and so legitimated as something we are naturally (and so, inevitably) impelled to. It is also one of the main reasons that have motivated the common misreading of Unamuno in pragmatist terms, as if Unamuno's religious faith were something we should voluntarily embrace after realizing its practical adequacy. For a detailed account of why we should not read Unamuno's "*hambre de inmortalidad*" as referring to the psychological, empirically contingent claim that we, human beings, desire for an endless existence, but to the stronger, metaphysical claim that *all* singular things (*i.e.*, not only human beings but also *prima facie* non-sentient beings such as plants and rocks) seek an endless existence, see Alberto Oya, *Unamuno's Religious Fictionalism* (Gewerbestrasse: Palgrave Macmillan, 2020), pp. 13–27. On why we should not consider Unamuno as a pragmatist philosopher in any philosophical relevant sense of the term, and why Unamuno's notion of religious faith cannot be identified with William James argument for religious belief as stated in his "The Will to Believe", see Alberto Oya, "Unamuno and James on Religious Faith" (*Teorema. Revista Internacional de Filosofía*, vol. XXXIX, n. 1 (2020), pp. 85–104).

any sort of justification for the claim that our natural appetite for an endless existence will be satisfied: the sort of immortality they attempt to demonstrate does not refer to *us*, the individuals we are here and now, it is not *our* immortality they are talking about. This point is repeatedly emphasized by Unamuno throughout most of his texts —take, for example, the following quote from his *Del sentimiento trágico de la vida en los hombres y los pueblos*:

> Without some kind of body or spirit-cover, the immortality of the pure soul is not true immortality. In the end, what we long for is a prolongation of this life, of this life and no other, this life of flesh and suffering, this life which we abominate at times precisely because it comes to an end.[5]

So, it seems that we cannot reasonably claim that our natural inclination for an endless existence will be satisfied. However, Unamuno argues, this conclusion seems avoidable if we are to accept the possibility of the biblical testimony about the Resurrection of all dead that is said to be announced (and exemplified) by Jesus Christ. According to the biblical testimony about the Resurrection of Jesus Christ, this kind of immortality is not restricted to some part of our human nature (*i.e.*, the human soul) but it refers to us, the "*hombres de carne y hueso*" that we are here and now —and therefore, in contrast with the traditional proofs for human immortality, this kind of immortality announced by Jesus Christ seems to succeed in preserving our own singularity. Furthermore, since Resurrection refers to an after earthly death existence, we can still hold to the reasonable claim that we are all going to die without this diminishing the possibility of enjoying an endless existence. So, Unamuno concludes, it seems that only if (the Christian) God exists, will we enjoy an endless existence. This is what allows Unamuno to shift the focus of his discourse from our natural appetite for an endless existence to our natural appetite for God — more concretely, Unamuno's argument here can be outlined as follows: we naturally seek an endless existence; only if (the Christian) God

[5] *The Tragic Sense of Life in Men and Nations*, p. 254 [*Del sentimiento trágico de la vida en los hombres y en los pueblos*, p. 236].

exists, will we enjoy an endless existence; therefore, we (*mediately*) seek God.[6]

Now, the problem is, according to Unamuno, that we are not justified, on an evidential basis, in claiming that God exists. Arguments from natural theology fail to demonstrate the existence of God because they start from the erroneous assumption that the existence of God can be inferred as being the only explanation (or, at least, the best explanation) for some worldly events. These arguments take the logical form of abductive inferences and, as such, they only work under the assumption that an explanation in terms of God's acting has some sort of explanatorily power. But theistic explanations, Unamuno says, have no explanatory power: God is not a scientific theoretical entity, and theism is not akin to a scientific hypothesis. God gives the world an ultimate meaning and purpose, but accepting the existence of God does not help us to explain why a given fact has occurred or why the world is such or such a way and not otherwise. God answers the "*¿para qué?*" ("wherefore?") of the world, but not its "*¿por qué?*" ("why?").[7] The claim that we cannot come to believe, on an evidential basis, that God exists is continuously present in the novel *San Manuel Bueno, mártir*, and it is what explains Manuel Bueno's inability to form the belief that God exists (and, hence, that he will enjoy an endless after earthly death existence). It also explains Lázaro's words to Ángela when he first meets Manuel Bueno:

[6] For a more detailed account of Unamuno's reasoning for claiming that as a consequence of the "*hambre de inmortalidad*" we all long for the Christian God and His Salvation, see Alberto Oya, *Unamuno's Religious Fictionalism* (Gewerbestrasse: Palgrave Macmillan, 2020), pp. 37–50.

[7] See *The Tragic Sense of Life in Men and Nations*, p. 168 [*Del sentimiento trágico de la vida en los hombres y en los pueblos*, p. 200]: "We need God, not in order to understand the *why*, but in order to feel and assert the ultimate *wherefore*, to give meaning to the Universe". For a more detailed account of Unamuno's reasoning for claiming that we cannot come to form on rational, evidential basis, the belief that the Christian God actually exists, see Alberto Oya, *Unamuno's Religious Fictionalism* (Gewerbestrasse: Palgrave Macmillan, 2020), pp. 40–43 and 51–57; see also Alberto Oya, "Unamuno and James on Religious Faith" (*Teorema. Revista Internacional de Filosofía*, vol. XXXIX, n. 1 (2020), pp. 85–104), pp. 95–98, and Alberto Oya, "Análisis de *Un pobre hombre rico o el sentimiento cómico de la vida*, de Miguel de Unamuno" (*Estudios Filosóficos*, vol. 70, n. 204 (2021), pp. 367–374).

> "Now this is something else again", he told me as soon as he came back from hearing Don Manuel for the first time. "He's not like the others; still, he doesn't fool me, he's too intelligent to believe everything he has to teach".[8]

In fact, Unamuno considered that any attempt to address the question of God in a rational way was ill-flawed from its very beginning. Take, for example, the so-called problem of evil. The obvious existence of evil and pain in the world seems to go against, or at least undermine, the core claim of theism that the world is the result of the intentional activity of an all-good and all-powerful supernatural being. At least with regard to natural evil, it seems that the only way to make the existence of evil consistent with the very notion of God (*i.e.*, as an all-good and all-powerful supernatural being) is by accepting our ignorance of God's intentions and purposes: although we cannot comprehend God's benevolence, we should rely on the assumption that God is an all-good being and so His actions are necessarily benevolent. But this is nothing more than recognizing our incapacity to comprehend God.[9] This point is nicely illustrated by Manuel Bueno's words:

> Often he [Manuel Bueno] used to accompany the doctor on his rounds, and stressed the importance of following the doctor's orders. Most of all he was interested in maternity cases and the care of children; it was his opinion that the old wives' sayings "from the cradle to heaven" and the other one about "little angels belong in heaven" were nothing short of blasphemy. The death of a child moved him deeply. "A stillborn child, or one who dies soon after birth are, like suicides, the most terrible mystery to me", I once heard him say, "Like a child crucified!"[10]

[8] *Saint Manuel Bueno, Martyr*, p. 153 [*San Manuel Bueno, mártir*, p. 1139].

[9] For a more developed characterization of this line of reasoning, see Alberto Oya, "Is it Reasonable to Believe that Miracles Occur?" (*Teorema. Revista Internacional de Filosofía*, vol. XXXVIII, n. 2 (2019), pp. 39–50).

[10] *Saint Manuel Bueno, Martyr*, p. 144 [*San Manuel Bueno, mártir*, p. 1134].

We are, then, not justified in believing that God exists. Shall we conclude from this that God does not exist? Unamuno's answer is in the negative: the lack of evidential justification for believing that God exists does not constitute positive evidence for forming the belief that God does not exist. And, Unamuno says, there is no argument which succeeds in demonstrating that God does not exist.[11] So, the most reasonable conclusion is to neither affirm nor deny the existence of God, but to accept that the question of God's existence is an open question which cannot be solved on an evidential, rational basis: "Reason does not prove to us that God exists, but neither does it prove that He cannot exist".[12]

Philosophical reasoning is, therefore, of no use here. Nonetheless, we cannot stop seeking God, in so far as we cannot silence our own nature, and only through God's grace will our most basic and natural inclination be satisfied. In such circumstances it is understandable that one might find desirable simply forgetting about epistemic justification and start believing without evidences, by a passional, irrational act of will, that God exists (and that He will concede us an endless existence). This attitude can be found in Manuel Bueno's longing for the faith he had when he was a child —and which is nothing more than what Unamuno in his philosophical essays called "*la fe del carbonero*" ("the faith of the charcoal burner"):[13]

> "Angelita, you have the same faith you had when you were ten, don't you? You believe, don't you?"
> "Yes, I believe, Father"
> "Then go on believing. And if doubts come to torment you,

[11] See Miguel de Unamuno, "My Religion", in *The Selected Works of Miguel de Unamuno (vol. 5)*, ed. and trans. Anthony Kerrigan (Princeton: Princeton University Press, 1974), p. 212 [Miguel de Unamuno, "Mi religión", in *Miguel de Unamuno: obras completas (vol. 3: "Nuevos ensayos")*, ed. Manuel García Blanco (Madrid: Escelicer, 1968 [1907]), p. 261]: "No one has succeeded in convincing me rationally of God's existence, but neither have they convinced me of His non-existence. The reasoning of atheists strikes me as being even more superficial and futile than that of their opponents".

[12] *The Tragic Sense of Life in Men and Nations*, p. 165 [*Del sentimiento trágico de la vida en los hombres y en los pueblos*, p. 198].

[13] *The Tragic Sense of Life in Men and Nations*, p. 84 [*Del sentimiento trágico de la vida en los hombres y en los pueblos*, p. 153].

> suppress them utterly, even to yourself. The main thing is to live..."[14]

Unamuno, however, explicitly rejects this possibility. We cannot willingly form the belief that God exists without being concerned about the evidence for the existence of God and simply because this belief being true is desirable to us. We cannot form the belief that God exists without committing ourselves to accepting the claim that God exists —such a thing would not be believing, but self-deception:

> The believer who resists examining the foundations of his belief is a man who lives in insincerity and lies. The man who does not want to think about certain eternal problems is a liar, nothing more than a liar.[15]

III. From the "*Sentimiento trágico de la vida*" to Religious Faith

So, we have seen that Unamuno is assuming the metaphysical claim that the most basic natural inclination of all singular things (not only sentient beings) is to seek an endless existence. Unamuno, of course, recognizes the obvious fact that we have overwhelming evidence to conclude that we all are going to die. Furthermore, Unamuno argues that traditional philosophical arguments for proving human immortality fail in their purpose because they do not succeed in preserving our own singularity, to continue being the same individuals of "*carne y hueso*" that we are here and now. Only if (the Christian) God exists, Unamuno says, will our appetite for an endless existence be satisfied —which would now refer to an after earthly death endless existence. The problem now is,

[14] *Saint Manuel Bueno, Martyr*, p. 160 [*San Manuel Bueno, mártir*, p. 1143].

[15] Miguel de Unamuno, "Verdad y vida", in *Miguel de Unamuno: obras completas (vol. 3: "Nuevos ensayos")*, ed. Manuel García Blanco (Madrid: Escelicer, 1968 [1908]), p. 266. My translation. The Spanish text reads: "El creyente que se resiste a examinar los fundamentos de su creencia es un hombre que vive en insinceridad y en mentira. El hombre que no quiere pensar en ciertos problemas eternos es un embustero, y nada más que un embustero".

however, that we lack any evidential support for believing either that God exists or that God does not exist. This is what Unamuno called the "*sentimiento trágico de la vida*" ("the tragic feeling of life"); *i.e.*, the struggle ("*agonía*") between our wanting an endless existence (and so, derivatively, our wanting God to exist) and our lack of evidential justification for believing that God exists (and so our lack of evidential justification for believing that we will enjoy an endless existence).

The "*sentimiento trágico de la vida*" ultimately arises as our reaction to the "*hambre de inmortalidad*", which is, according to Unamuno, our most basic and natural inclination —this is why Unamuno calls it "*de la vida*" ("of life"). Likewise, since our longing for an endless existence is a natural, non-intellectual need, something we are impelled to because of our own nature, the "*sentimiento trágico de la vida*" is not a theoretical struggle but a sentimental one, something we intimately feel —this is why Unamuno calls it "*sentimiento*" ("feeling").[16] The conflict is "*trágico*" ("tragic") because it is irresoluble: we cannot override our lack of evidential justification by voluntarily forming the belief that God exists (or that God does not exist) because our beliefs aim at truth (*i.e.*, we cannot believe that P without believing that P is true), and neither can we suspend our judgment and resign ourselves to doubt since this would amount to silencing our most basic natural inclination. As Manuel Bueno says, once we become aware of our situation, there is no turning back:

> Like Moses, I have seen the face of God —our supreme dream— face to face, and as you already know, and as the Scriptures say, he who sees God's face, he who sees the eyes

[16] See *The Tragic Sense of Life in Men and Nations*, pp. 121–123 [*Del sentimiento trágico de la vida en los hombres y en los pueblos*, pp. 174–175]: "The question of the immortality of the soul, of the persistence of individual consciousness, is not a rational concern, it falls outside the scope of reason. As a problem —whatever solution is assumed— it is irrational. Rationally, even the stating of the problem lacks of sense. The immortality of the soul is as inconceivable as, strictly speaking, its absolute mortality would be. For purposes of explaining the world and existence —and such is the task of reason— there is no need to suppose that our soul is either mortal or immortal. The very statement of the supposed problem then, is irrational. [...] This vital longing is not properly speaking a problem, it can not be given any logical status, it can not be formulated in propositions rationally disputable; but it poses itself as a problem the way hunger poses itself as a problem".

> of the dream, the eyes with which He looks at us, will die inexorably and forever.[17]

Our incapacity to solve the struggle, Unamuno says, causes us a sort of anguish. This spiritual suffering, however, despite being painful and inescapable, should not lead us to refusing to enjoy this earthly life. Again, our longing for an endless existence is our most basic and natural inclination, which means that we cannot stop desiring to exist. This attitude is clearly present in Manuel Bueno, when he exclaims: "Yes! One must live".[18] And this is what explains why Manuel Bueno, who fully embodies the spiritual suffering that the "*sentimiento trágico de la vida*" carries with it, and who at times goes on to define his own life as a "[...] kind of continual suicide, or a struggle against suicide [...]",[19] never stops valuing his earthly life. In fact, Manuel Bueno does not hesitate in claiming that the lack of desire to enjoy life is "a thousand time worse than hunger":

> Listen, Lázaro, I have helped poor villagers to die well, ignorant, illiterate villagers who had scarcely ever been out of their village, and I have learned from their own lips, or sensed it when they were silent, the real cause of their sickness unto death, and there at their deathbed I have been able to see into the black abyss of their life —weariness. A weariness a thousand time worse than hunger![20]

[17] *Saint Manuel Bueno, Martyr*, p. 170 [*San Manuel Bueno, mártir*, p. 1148].

[18] *Saint Manuel Bueno, Martyr*, p. 161 [*San Manuel Bueno, mártir*, p. 1144].

[19] *Saint Manuel Bueno, Martyr*, p. 163 [*San Manuel Bueno, mártir*, p. 1144].

[20] *Saint Manuel Bueno, Martyr*, p. 163 [*San Manuel Bueno, mártir*, p. 1144]. See also *Saint Manuel Bueno, Martyr*, p. 145 [*San Manuel Bueno, mártir*, p. 1134]: "'The most important thing', he [Manuel Bueno] would say, 'is for the people to be happy; everyone must be happy just to be alive. To be satisfied with life is of first importance. No one should want to die until it is God's will'. [...] Once he commented at a wedding: 'Ah, if I could only change all the water in our lake into wine, into a gentle little wine which, no matter how much of it one drank, would always make one joyful without making one drunk... or, if it made one drunk, would make one joyfully tispy'".

Actually, this spiritual suffering is not something we must try to silence or avoid, but is rather a "*dolor sabroso*" ("sweet-tasting pain").[21] The anguishing situation that results from our incapacity to escape from the "*sentimiento trágico de la vida*" is something desirable in itself since it is precisely from this spiritual suffering that religious faith emerges:

> All these speculative confessions amount to so much wretchedness, I know; but from the depths of wretchedness springs new life, and it is only by draining the dregs of spiritual sorrow that the honey at the bottom of life's cup is tasted. Anguish leads us to consolation.[22]

By suffering, Unamuno says, we become aware of our miserable and tragic situation, faced with which we can do nothing but commiserate with ourselves. Thus, our spiritual suffering makes way for compassion. And compassion is where love originates, since when we commiserate with someone we are also loving them: we only worry for those we take into consideration. But, according to Unamuno, we are not alone in this suffering. As soon as we realize of the universality of the "*hambre de inmortalidad*", that the longing for an endless existence is the most basic and natural inclination of all singular things, we come to realize that the entire world shares our anguishing condition with us. This allows us to comprehend Manuel Bueno's suffering when contemplating the lake —there, alone with nature, is where he realizes the universality of the "*sentimiento trágico de la vida*":

> "What an incredible man!" he [Lázaro] exclaimed to me [Ángela] once. Yesterday, as we were walking along beside

[21] See *The Tragic Sense of Life in Men and Nations*, p. 307 [*Del sentimiento trágico de la vida en los hombres y en los pueblos*, p. 275]: "There is no point in taking opium; it is better to put salt and vinegar in the soul's wound, for if you fall asleep and no longer feel the pain, then you no longer exist. And the point is to exist. Do not, then, close your eyes before the overawing Sphinx, but gaze on her face to face, and let her take you in her mouth and chew you with her hundred thousand poisonous teeth and swallow you up. And when she has swallowed you, you will know the sweet taste of suffering".

[22] *The Tragic Sense of Life in Men and Nations*, p. 64 [*Del sentimiento trágico de la vida en los hombres y en los pueblos*, p. 143].

> the lake he [Manuel Bueno] said: "There lies my greatest temptation." [...] "How that water beckons me with its deep quiet!... an apparent serenity reflecting the sky like a mirror —and beneath it the hidden current! [...]." And then he added: "Here the river eddies to form a lake, so that later, flowing down the plateau, it may form cascades, waterfalls and torrents, hurling itself through gorges and chasms. Thus life eddies in the village; and the temptation to commit suicide is greater beside the still waters which at night reflect the starts, than it is beside the crashing falls which drive one back in fear."[23]

Since only conscious, living beings suffer, claiming that the whole world suffers as we do is tantamount to adopting a religious understanding of the world —*i.e.*, we cease seeing the world as an *it* and start seeing it *as if it were* a conscious, personal living Being. By becoming aware of the universality of our anguishing situation, we come to commiserate with and love the whole world. And compassion, in its practical, ethical sense takes the form of charity. To cultivate charity, Unamuno says, is to act in such a way as to lovingly give ourselves over to the spiritual care of others. Charity is an attempt to liberate ourselves and the entire world from the spiritual suffering and the tragic situation in which we all live: it is through the practice of charity that we *come to feel as* part of others and so we somehow surpass our own individuality without ceasing to be the individuals of "*carne y hueso*" we are here and now. And by the practice of charity, by our agapeic giving ourselves to the world and leaving our mark on it, we come to spiritualize the world —which is tantamount to saying that *we feel as if there were* some sort of communion between us and the world as a Conscience, as God. We find exemplified this feeling of communion with the world in Ángela's words at the end of the novel:

[23] *Saint Manuel Bueno, Martyr*, pp. 162–163 [*San Manuel Bueno, mártir*, p. 1144]. See also Ángela's words to Lázaro after Manuel Bueno's death: "'Don't stare into the lake so much', I begged him" (*Saint Manuel Bueno, Martyr*, p. 174 [*San Manuel Bueno, mártir*, p. 1150]).

> One must live! And he [Manuel Bueno] taught me to live, he taught us to live, to feel life, to feel the meaning of life, to merge with the soul of the mountain, with the soul of the lake, with the soul of the village, to lose ourselves in them so as to remain in them forever. He taught me by his life to lose myself in the life of the people of my village, and I no longer felt the passing of the hours, and the days, and the years, any more than I felt the passage of the water in the lake. It began to seem that my life would always be like this. I no longer felt myself growing old. I no longer lived in myself, but in my people, and my people lived in me.[24]

Ultimately, Unamuno's point is that carrying out an agapeic way of life, commiserating with and lovingly giving oneself to the whole world, does not constitute a diminishment of one's own singularity but is rather the only way to increase it. It is only through the agapeic giving of ourselves that we come to feel in communion with the whole world while preserving our own singularity, while continuing to be the same individuals we are here and now. According to Unamuno, then, an agapeic way of life is not merely consistent with human nature but an affirmation of it, the expression of our natural and most basic inclination to increase our own singularity.[25]

[24] *Saint Manuel Bueno, Martyr*, p. 176 [*San Manuel Bueno, mártir*, p. 1152]. I have modified Kerrigan's translation of the first sentence of this quote. The original Spanish text reads: "¡Hay que vivir!". Kerrigan translates it as "Life must go on!". A more accurate translation of this sentence is: "One must live!".

[25] This is what explains Unamuno's comments on Nietzsche's criticisms of the Christian, agapeic way of life (see, *e.g.*, Miguel de Unamuno, "Uebermench", in *Miguel de Unamuno: obras completas (vol. IV: "La raza y la lengua")*, ed. Manuel García Blanco (Madrid: Escelicer, 1966 [1914], pp. 1367-1369). Unamuno's reasoning starts from a metaphysical assumption similar to that of Nietzsche—that is, a modified version of Spinoza's *conatus*, construed not only in terms of self-preservation but also in terms of increase of power. Whereas Nietzsche claimed that a Christian, agapeic way of life is something antinatural insofar as it goes against the natural tendency to increase one's own power, Unamuno responded by arguing that an agapeic way of life is precisely a direct consequence of this natural tendency. It is through our agapistic giving of ourselves over to the whole world that we come to *feel communed* with the entire world, and so we *somehow* come to surpass our own individuality without losing our own personal identity, without ceasing to

Now it becomes evident why Manuel Bueno, whose religion consists in "[...] consoling myself by consoling others, even though the consolation I give them is not ever mine [...]",[26] embodies Unamuno's conception of religious faith. It is precisely *because* Manuel Bueno is unable to rid himself of his doubts and fails to come to believe that God exists, that he devotes himself to the spiritual care of his people. If there is something that defines Manuel Bueno it is his selfless giving to the care of others:

> How he loved his people! He spent his life salvaging wrecked marriages, forcing unruly children to submit to their parents, or reconciling parents to their children, and, above all, he consoled the embittered and weary in spirit and helped everyone to die well.[27]

Religious faith, then, is expressed in the practice of charity, which is a practical, non-theoretical issue. This is what explains why Manuel Bueno does not like engaging in theological discussions: religious faith is to give oneself to the others, not to save oneself by getting lost in intricate theological thoughts.

> His [Manuel Bueno's] life was active rather than contemplative, and he constantly fled from idleness, even

be the individuals of "*carne y hueso*" we are here and now. And once conceded that what emerges from this natural tendency to increase one's own power is Unamuno's notion of religious faith, then Nietzsche's ideal of the Overman, and his implied denial of the Christian understanding of the world, is nothing more than a cowardly self-deception, an attempt to silence one's own natural anguished condition (*i.e.*, the "*sentimiento trágico de la vida*") instead of accepting it by making it the foundation of his acting and understanding of the world (*i.e.,* Unamuno's religious faith). For a more detailed account on why Unamuno's defense of religious faith can be read as a response to Nietzsche's criticisms of the Christian, agapeic way of life, see Alberto Oya, "Nietzsche and Unamuno on *Conatus* and the Agapeic Way of Life" (*Metaphilosophy*, vol. 51, nos. 2–3 (2020), pp. 303–317).

[26] *Saint Manuel Bueno, Martyr*, p. 159 [*San Manuel Bueno, mártir*, p. 1142].

[27] *Saint Manuel Bueno, Martyr*, p. 138 [*San Manuel Bueno, mártir*, p. 1131]. See also *Saint Manuel Bueno, Martyr*, p. 139 [*San Manuel Bueno, mártir*, p. 1131]: "He treated everyone with the greatest kindness; if he favored anyone, it was the most unfortunate, and especially those who rebelled".

from leisure. Whenever he heard it said that idleness was the mother of all vices, he added: "And also of the greatest vice of them all, which is to think idly". Once I asked what he meant and he answered: "Thinking idly is thinking as a substitute for doing, or thinking too much about what is already done instead of about what must be done. What's done is done and over with, and one must go on to something else, for there is nothing worse than remorse without possible solution" Action! Action! [...] And so it was that he was always busy, sometimes even busy looking for things to do. He wrote very little on his own, so that he scarcely left us anything in writing, not even notes; on the other hand, he acted as scribe for everyone else, especially composing letters for mothers to their absent children. He also worked with his hand, pitching in to help with some of the village tasks. At threshing time he reported to the threshing floor to flair and winnow, meanwhile teaching and entertaining the workers by turn. Sometimes he took the place of a worker who had fallen sick. One bitter winter's day he came upon a child half-dead with cold. The child's father had sent him into the woods to bring back a calf that had strayed. "Listen", he said to the child, "you go home and get warm, and tell your father that I am bringing back the calf". [...] In winter he chopped wood for the poor. [...] He also was in the habit of making handballs for the boys and many toys for the younger children. [...] Often he would visit the local school too, to help the teacher, to teach alongside him —and not only the catechism. The simple truth was that he fled relentlessly from idleness and from solitude. He went so far in this desire of his to mingle with the villagers, especially the young people and the children, that he event attended the village dances. And more than once he played the drum to keep time for the boys and girls dancing; this kind of activity, which in another priest would have seemed like

> a grotesque mockery of his calling, in him somehow took on the appearance of a divine office.[28]

We have just seen that Manuel Bueno embodies Unamuno's conception of religious faith. However, it is worth mentioning that Manuel Bueno is not the only character in the novel in whom we can find expressed Unamuno's conception of religious faith. We also find it exemplified, albeit perhaps in a subtler way, in the clown who continues to work and make others laugh despite his wife being mortally ill. Why does Manuel Bueno not hesitate in calling the clown a "Saint"? It is because his actions are not (at least, not exclusively) driven by a selfish motivation, but by the purpose of taking care of others and making their life more enjoyable:

> One day a band of poor circus people came through the village. Their leader —who arrived with a gravely ill and pregnant wife and three children to help him— played the clown. While he was in the village square making all the children, and even some of the adults, laugh with glee, his wife suddenly fell desperately ill and had to leave; she went off accompanied by a look of anguish from the clown and a howl of laughter from the children. Don Manuel hurried after her, and a little later, in a corner of the inn's stable, he helped her give up her soul in a state of grace. When the performance was over and the villagers and the clown learned of the tragedy, they came to the inn, and there the poor, bereaved clown, in a voice overcome with tears, said to Don Manuel, as he took his hand and kissed it: "They are quite right, Father, when they say you are a saint". Don Manuel took the clown's hand in his and replied in front

[28] *Saint Manuel Bueno, Martyr*, pp. 143–144 [*San Manuel Bueno, mártir*, p. 1133–1134]. See also *Saint Manuel Bueno, Martyr*, p. 147 [*San Manuel Bueno, mártir*, p. 1135]: "It is not at all because my sister is a widow and I have her children and herself to support —for God looks after the poor— but rather because I simply was not born to be a hermit, an anchorite; the solitude would crash my soul; and, as far as a monastery is concerned, my monastery is Valverde de Lucerna. I was not meant to live alone, or die alone. I was meant to live for my village, and die for it too. How should I save my soul if I were not to save the soul of my village as well?"

> of everyone: "It is you who are the saint, good clown. I watched you at your work and understood that you do it not only to provide bread for your own children, but also to give joy to the children of others. And I tell you now that your wife, the mother of your children, whom I sent to God while you worked to give joy, is at rest in the Lord, and that you will join her there, and that the angels, whom you will make laugh with happiness in heaven, will reward you with their laughter".[29]

IV. Religious Faith is not Believing

We have just seen that, according to Unamuno, it is precisely from doubt, from our lack of evidential support for believing neither that God exists nor that God does not exist, together with our *natural* appetite for an endless existence, that religious faith emerges. Doubt, therefore, is essential to religious faith: without doubt, there is no faith. Thus, in Unamuno's schema, incredulity is not an impediment, but the cause of a holy life. Something I must emphasize here is that Unamuno's religious faith does not aim to put an end to doubt: the "*sentimiento trágico de la vida*" remains tragic, irresoluble, no matter what we do. Unamuno's religious faith consists in adopting a religious understanding of the world and in entering into a sort of personal relationship with it. But this religious understanding of the world is not a description of the world, it does not lead us to form the belief that God exists or that He will bless us with an endless after earthly death existence. And this is so because Unamuno's faith is justified as being a consequence of our own human nature (*i.e.*, something we are naturally, and so inevitably, led to), not because of its being true.[30] That religious faith does not solve the question

[29] *Saint Manuel Bueno, Martyr*, pp. 145–146 [*San Manuel Bueno, mártir*, p. 1134–1135].

[30] For a more detailed account on why the kind of religious understanding of the world Unamuno's religious faith consists in is not a description of how the world actually is, see Alberto Oya, *Unamuno's Religious Fictionalism* (Gewerbestrasse: Palgrave Macmillan, 2020), especially pp. 59–86.

of God's existence is nicely illustrated in the following conversation between Lázaro and Manuel Bueno:

> When I [Lázaro] said to him: "Is it really you, the priest, who suggests that I pretend?" he [Manuel Bueno] replied, hesitatingly: "Pretend? Not at all! It would not be pretending. 'Dip your fingers in holy water, and you will end by believing', as someone said". And I, gazing into his eyes, asked him: "And you, by celebrating the Mass, have you ended up by believing?" He looked away and stared out the lake, until his eyes filled with tears. And it was in this way that I came to understand his secret.[31]

Again, Unamuno's religious faith is not a theoretical, intellectual issue. Religious faith is not believing, it does not consist in accepting as a truth that the world is such and such and not otherwise. Religious faith is nothing more than our subjective, non-evidentially grounded but experientially felt, understanding of the world. And this religious understanding of the world, in its practical, ethical sense, is expressed through the practice of charity: in a loving, agapeic giving to the whole world. This is precisely what Manuel Bueno means when he says that:

> As for true religion, all religions are true insofar as they give spiritual life to the people who profess them, insofar as they console them for having been born only to die. And for each race the truest religion is their own, the religion that made them... And mine? Mine consists in consoling myself by consoling others, even though the consolation I give them is not ever mine.[32]

This non-theoretical nature of religious faith is present in Manuel Bueno: his already commented refusal to enter into theological disquisitions

[31] *Saint Manuel Bueno, Martyr*, p. 157 [*San Manuel Bueno, mártir*, p. 1141].

[32] *Saint Manuel Bueno, Martyr*, pp. 158–159 [*San Manuel Bueno, mártir*, p. 1142].

should be read in this context.[33] It is also what explains that there are no intellectual motives behind Lázaro's conversion.[34] This non-theoretical nature of religious faith also allows us to understand the role of the character Blasillo, who receives the appellative "*el bobo*" ("the fool") because of his lack of intellectual development. Blasillo accompanies Manuel Bueno in delivering his masses and he continuously repeats, presumably without understanding its meaning and simply as an act of imitation, the words of Jesus Christ that Manuel Bueno so vividly exclaims in his masses: "My God, my God, why hast Thou forsaken

[33] See *Saint Manuel Bueno, Martyr*, p. 149 [*San Manuel Bueno, mártir*, p. 1137]: "Another time in the confessional I told him [to Manuel Bueno] of a doubt which assailed me, and he responded: 'As to that, you know what the catechism says. Don't question me about it, for I am ignorant; in Holy Mother Church there are learned doctors of theology who will know how to answer you'". See also *Saint Manuel Bueno, Martyr*, p. 164 [*San Manuel Bueno, mártir*, p. 1145]: "Don Manuel had to moderate and temper my brother's zeal and his neophyte's rawness. As soon as he heard that Lázaro was going about inveighing against some of the popular superstitions he told him firmly: 'Leave them alone! It's difficult enough making them understand where orthodox belief leaves off and where superstition begins. And it's even harder for us. Leave them alone, then, as long as they get some comfort... It's better for them to believe everything even things that contradict one another, that to believe nothing. The idea that someone who believes too much ends up not believing anything is a Protestant notion. Let us not protest! Protestation destroys contentment and peace'".

[34] See *Saint Manuel Bueno, Martyr*, p. 156–157 [*San Manuel Bueno, mártir*, p. 1141]: "'Lázaro, Lázaro, what joy you have given us all today; the entire village, the living and the dead, especially our mother. Did you see how Don Manuel wept for joy? What joy you have given us all!' / 'That's why I did it', he answered me. / 'Is that why? Just to give us pleasure? Surely you did it for your own sake, because you were converted'. / [...] Thereupon, serenely and tranquilly, in a subdued voice, he recounted a tale that cast me into a lake of sorrow. He told me how Don Manuel had begged him, particularly during the walks to the ruins of the old Cistercian abbey, to set a good example, to avoid scandalizing the townspeople, to take part in the religious life of the community, to feign belief even if he did not feel any, to conceal his own ideas —all this without attempting in any way to catechize him, to instruct him in religion, or to effect a true conversion." Notice that Kerrigan's translation of this last sentence is inaccurate. The sentence "to instruct him in religion, or to effect a true conversion" does not appear in Unamuno's text. The original Spanish text reads: "[...] para que ocultase sus ideas al respecto, mas sin intentar siquiera catequizarle, convertirle de otra manera". A more accurate translation of this sentence is: "[...] to conceal his own ideas, without even trying to catechize him, convert him in a different way".

me?" (Psalms 22: 1).[35] Blasillo dies at the same time and in a similar way to Manuel Bueno:

> He [Manuel Bueno] was carried to the church and taken, in his armchair, into the chancel, to the foot of the altar. In his hand he held a crucifix. My brother and I stood close to him, but the fool Blasillo wanted to stand even closer. He wanted to grasp Don Manuel by the hand, so that he could kiss it. When some of the people nearby tried to stop him, Don Manuel rebuked them and said:
> "Let him come closer... Come, Blasillo, give me your hand"
> The fool cried for joy. And then Don Manuel spoke [...]. Then he gave his blessing to the whole village, with the crucifix held in his hand, while the women and children cried and even some of the men wept softly. Almost at once the prayers were begun. Don Manuel listened to them in silence, his hand in the hand of Blasillo the fool, who was falling asleep to the sound of the praying. [...] On reaching "The Resurrection of the flesh and life everlasting" the people sensed that their saint had yielded up his soul to God. It was not necessary to close his eyes even, for he died with them closed. When we tried to wake up Blasillo, we found that he, too, had fallen asleep in the Lord forever. So that later there were two bodies to be buried.[36]

That "*el bobo*" died in similar circumstance to a Saint, and especially the fact that at the end of the novel Blasillo is no longer qualified as "*el bobo*" but as "a Saint", illustrates Unamuno's claim that no special intellectual faculty is needed to exercise holiness.[37]

[35] See *Saint Manuel Bueno, Martyr*, p. 140 [*San Manuel Bueno, mártir*, p. 1132]: "Afterwards the fool Blasillo went about piteously repeating, like an echo, 'My God, my God, why hast Thou forsaken me?' with such effect that everyone who heard him was moved to tears, to the great satisfaction of the fool, who prided himself on this triumph of imitation".

[36] *Saint Manuel Bueno, Martyr*, pp. 171–172 [*San Manuel Bueno, mártir*, p. 1149].

[37] See *San Manuel Bueno, mártir*, p. 1152: "[...] also about the memory of the poor Blasillo, my Saint Blasillo, and may he take care of me from heaven". My

V. Unamuno's Religious Faith as a Return to Early Christianity

As I have already said, Unamuno's defense of religious faith depends on accepting *the possibility* of the biblical testimony of the Resurrection of Jesus Christ: if the Resurrection of Jesus Christ were something impossible (if, for example, the possibility of God's intervening in the natural world were something ruled out a priori, as being inconsistent with the very notion of God), or if the sort of immortality promised by the Christian God did not succeed in preserving our own individuality, the "*sentimiento trágico de la vida*" would never arise because there would be no connection between the existence of God and the satisfaction of our natural inclination for an endless existence. It is important to emphasize, however, that Unamuno is not assuming that the Resurrection really did occur; what Unamuno's argument assumes is that the Resurrection of Jesus Christ, despite being something which cannot be solved on an evidential basis, is an open possibility. If we were justified in accepting that Jesus Christ resurrected, then this very belief would give us evidence for the belief that God exists but, as we have just seen, the "*sentimiento trágico de la vida*" depends on accepting our lack of evidential support for believing that God exists.

Unamuno was, of course, well aware that his conception of religious faith did not fit with any conventional understanding of Christianity, but in so far as all his reasoning depends on the acceptance of the possibility of the Resurrection of Jesus Christ, Unamuno was somehow right in considering himself as a Christian. According to Unamuno, however, his position was not a reformulation of how Christian religious faith should be understood, but a return to the authentic, original conception of Christianity. His continuous references to the Bible, and especially to the "My God, my God, why hast Thou forsaken me?" (Psalms 22: 1) and "Lord, I believe, help thou mine unbelief!" (Mark 9: 24), should be read in this context.

translation. The Spanish text reads: "[...] y también sobre la memoria del pobre Blasillo, de mi san Blasillo, y que él me ampare desde el cielo". Notice that Kerrigan translates this sentence as: "[...] and even on the memory of the poor fool Blasillo, my Saint Blasillo —and may he help me in heaven!" (*Saint Manuel Bueno, Martyr*, p. 177). Kerrigan's translation is here inaccurate, since there is no "*bobo*" ("fool") in Unamuno's text.

Unamuno's claim that he was arguing for a return to the original conception of Christianity is present in most of his texts, and can be found as early as 1897, in his "¡Pistis y no gnosis!", where he explicitly claimed that for early Christians, faith was not to believe that God exists (*gnosis*) but to hope for God's Salvation (*pistis*).

> The youth of the Christian communities awaited the next coming of the kingdom of the Son of God; the person and the life of the Divine Master were the compass of their yearnings and feelings. They felt swelled with real faith, with what is confused with hope, with what is called *pistis*, faith or trust, religious faith not theological faith, pure faith that is still free of dogmas. They lived a life of faith; they lived for faith in the future; waiting for the kingdom of eternal life, they lived life. […] As the heat of faith dissipated and religion became more worldly, […] the juvenile *pistis* was substituted by gnosis, knowledge; belief, not strictly faith; doctrine, not hope. Believing is not trusting. Faith became the adhesion of the intellect; what knowledge of life is began to be taught; converting the aims of religious practices into philosophical, theoretical principles, and religion into metaphysics revealed. Sects, schools, dissents, dogmas were finally born. […] From then on, faith for many Christians was believing what we cannot see, *gnosis*, and not trusting in the kingdom of eternal life, *pistis*, in other words, believing what we did not see.[38]

[38] Miguel de Unamuno, "¡Pistis y no Gnosis!", in *Miguel de Unamuno: obras completas (vol. 3: "Nuevos ensayos")*, ed. Manuel García Blanco (Madrid: Escelicer, 1968 [1897]), pp. 682–683. My translation. The Spanish text reads: "Jóvenes las comunidades cristianas, esperaban la próxima venida del reino del Hijo de Dios; la persona y la vida del Divino Maestro eran el norte de sus anhelos y sentires. Sentíanse henchidas de verdadera fe, de la que con esperanza se confunde, de lo que se llamó *pistis*, fe o confianza, fe religiosa y no teologal, fe pura y libre todavía de dogmas. Vivían vida de fe; vivían por la esperanza en el porvenir; esperando el reino de la vida eterna, vivían ésta. [...] A medida que el calor de la fe iba menguando y mundanizándose la religión, [...] [l]a juvenil *pistis* fue siendo sustituída por la *gnosis*, el conocimiento; la creencia, y no propiamente la fe; la doctrina y no la esperanza. Creer no es confiar. Hízose de la fe adhesión del intelecto; empezóse a enseñar qué es el conocimiento de la vida; convirtiéronse los fines prácticos religiosos en

In his *La agonía del cristianismo* (1924) [*The Agony of Christianity*], Unamuno did not hesitate in claiming that theological dogmas appear with Saint Paul, not with Jesus Christ: "St. Paul made the Gospel biblical, changing the Word into the Letter".[39] The Church, Unamuno says, aimed to silence all doubts regarding the question of God's existence by dogmatically affirming the *truth* of Christianity. But this is nothing more than removing the "*sentimiento trágico de la vida*" and, with it, the very essence of Christianity: without doubt there is no Christian faith.

In fact, once Unamuno's conception of religious faith is accepted, the very idea of a Christian Church seems to be off the point. As we have seen, according to Unamuno, religious faith is something we intimately feel.[40] Religious faith, therefore, has nothing to do with, and should not be confused with, politics, economics or any other social issues. This explains Manuel Bueno's refusal to form an agrarian syndicate[41] and to

principios teóricos filosóficos, la religión en metafísica revelada. Nacieron sectas, escuelas, disidencias, dogmas por fin. [...] En adelante la fe fue para muchos cristianos creer lo que no vimos, *gnosis*, y no confiar en el reino de la vida eterna, *pistis,* es decir, creer lo que no vemos".

[39] Miguel de Unamuno, *The Agony of Christianity*, in *The Selected Works of Miguel de Unamuno (vol. 7)*, ed. and trans. Anthony Kerrigan (Princeton: Princeton University Press, 1974), p. 27 [Miguel de Unamuno, *La agonía del Cristianismo*, in *Miguel de Unamuno: obras completas (vol. 7: "Meditaciones y ensayos espirituales)*, ed. Manuel García Blanco (Madrid: Escelicer, 1966 [1924]), p. 320], Notice that this distinction between word and letter is the same distinction we found between theology and religion at the end of *San Manuel Bueno, mártir*: "The poor priest who came to replace Don Manuel found himself overwhelmed in Valverde de Lucerna by the memory of the saint, and he put himself in the hands of my brother and myself for guidance. He wanted only to follow in the footsteps of the saint. And my brother told him: 'Very little theology, Father, very little theology. Religion, religion, religion'. Listening to him, I smiled out myself, wondering if this were not a kind of theology too." (*Saint Manuel Bueno, Martyr*, pp. 173–174 [*San Manuel Bueno, mártir*, p. 1150]).

[40] See Miguel de Unamuno, *The Agony of Christianity*, p. 5 [Miguel de Unamuno, *La agonía del Cristianismo*, p. 308]): "Though, in actual fact, is there any Christianity outside each one of us?".

[41] See *Saint Manuel Bueno, Martyr*, p. 165 [*San Manuel Bueno, mártir*, pp. 1145–1146]): "'A syndicate?' Don Manuel replied sadly. 'A syndicate? And what is that? The Church is the only syndicate I know of. And you have certainly heard 'My kingdom is not of this world'. Our kingdom, Lázaro, is not of this world...' / 'And of the other?' / Don Manuel bowed his head: 'The other is here. Two kingdoms exist in this world. Or rather, the other world... Ah, I don't really

aid civil justice.[42] Having said this, it is also true that Unamuno did not reject the possibility of a Church in itself, but he rejected the notion of a Church as a sociopolitical institution.[43] Unamuno nowhere denies the legitimacy of a Church understood as a congregation of religious men. As

know what I am saying. But as for the syndicate, that's a carry-over from your radical days. No, Lázaro, no; religion does not exist to resolve the economic or political conflicts of this world, which God handed over to men for their disputes. Let men think and act as they will, let them console themselves for having been born, let them live as happily as possible in the illusion that all this has a purpose. I don't purpose to advise the poor to submit to the rich, nor to suggest to the rich that they submit to the poor; but rather to preach resignation in everyone, and charity toward everyone. For even the rich man must resign himself —to his riches, and to life; and the poor man must show charity —even to the rich. The Social Question? Ignore it, for it is none of our business. [...] No, Lázaro, no; no syndicates for us. If *they* organize them, well and good —they would be distracting themselves in that way. Let them play at syndicates, if that makes them happy'."

[42] See *Saint Manuel Bueno, Martyr*, pp. 140–141 [*San Manuel Bueno, mártir*, p. 1132]): "The priest's effect on people was such that no one ever dared to tell him a lie, and everyone confessed to him without need of a confessional. So true was this that one day, after a revolting crime had been committed in a neighboring village, the judge —a dull fellow who badly misunderstood Don Manuel— called on the priest and said: / 'Let's see if *you*, Don Manuel, can get this bandit to admit the truth'. / 'So that *you* may punish him afterwards?' asked the saintly man. 'No, judge, no; I will not extract from any man a truth which could be the death of him. That is a matter between him and his God... Human justice is none of my affair. 'Judge not that ye be not judged', said Our Lord'. / 'But the fact is, Father, that I, a judge...' / 'I understand. You, judge, must render unto Caesar that which is Caesar's, while I shall render unto God that which is God's'. / And, as Don Manuel departed, he gazed at the suspected criminal and said: 'Make sure, only, that God forgives you, for that is all that matters'."

[43] See Miguel de Unamuno, "Religión y patria", in *Miguel de Unamuno: obras completas (vol. 1: "Paisajes y ensayos")*, ed. Manuel García Blanco (Madrid: Escelicer, 1966 [1904]), pp. 1110–1111: "[...] the Catholic Church was not instituted to promote culture, but to save souls. […] Neither the Catholic Church was instituted to promote culture, nor were religious orders they founded designed to make or break homelands; the Church itself must have nothing to do with disputes between princes and states. The alliance between the Altar and the Throne is, in the long term, deadly for both.". My translation. The Spanish text reads: "[...] la Iglesia católica no se instituyó para promover la cultura, sino para salvar las almas. [...] Ni la Iglesia católica se instituyó para promover la cultura, ni las Órdenes religiosas que de ella han nacido tienen por misión hacer ni deshacer patrias, ni la Iglesia misma debe tener que ver con disputas de príncipes y de Estados. La alianza entre el Altar y el Trono es, a la larga, fatal a uno y a otro".

Manuel Bueno says, the sense of the Church should be "*bien entendido*" ("well understood").[44]

Throughout his entire *San Manuel Bueno, mártir*, we find Unamuno's claim that his conception of religious faith is a return to the original meaning of Christianity. To start with, "Manuel" is the Spanish name for "Immanuel", a Hebrew word meaning "God with us" —and "Bueno" is the Spanish word for "Good". Similarly, the narrator of the novel receives the name of "Ángela", which is derived from the Latin word "angelus", meaning a "messenger". And Unamuno's Lázaro, like the biblical Lazarus who the Christian Scriptures say was raised from the dead by Jesus Christ (John, 11: 43–44), is raised from his spiritual drowsiness by Manuel Bueno:

> "It was he [Manuel Bueno]", said my brother, "who made me into a new man. I was a true Lazarus whom he raised from the dead. He gave me faith".[45]

That Unamuno was aiming to draw a parallel between Manuel Bueno and Jesus Christ is already made explicit right at the very beginning of the novel, when Manuel Bueno is confused with Jesus Christ:

> And when on Good Friday he chanted, "My God, my God, why hast Thou forsaken me?" a profound shudder swept through the multitude, like the lash of the northeast wind across the waters of the lake. It was as if these people heard

[44] In the original Spanish text, the quote I am referring to here reads as follows: "Y tú, Lázaro, cuando hayas de morir, muere como yo, como morirá nuestra Ángela, en el seno de la Santa Madre Católica Apostólica Romana, de la Santa Madre Iglesia de Valverde de Lucerna, bien entendido" (*San Manuel Bueno, mártir*, p. 1148). Kerrigan translates it as: "And Lázaro, when your hour comes, die as I die, as Ángela will die, in the arms of the Holy Mother Church, Catholic Apostolic, and Roman; that is to say, the Holy Mother Church of Valverde de Lucerna" (*Saint Manuel Bueno, Martyr*, p. 169). Kerrigan's translation is here inaccurate since it simply forgets translating the "*bien entendio*" ("well understood"), which is, I think, the interesting point of the quote. A more accurate translation is: "And you, Lázaro, when you should die, die as I die, as our Ángela will die, in the arms of the Holy Roman Catholic Aposthოlic Mother, the Holy Mother Church of Valverde de Lucerna, well understood".

[45] *Saint Manuel Bueno, Martyr*, p. 173 [*San Manuel Bueno, mártir*, p. 1150].

> Our Lord Jesus Christ Himself, as if the voice sprang from the ancient crucifix, at the foot of which generations of mothers had offered up their sorrows.[46]

Unamuno's advocating for what he takes to be the original meaning of the message conveyed by Jesus Christ helps us to comprehend why Manuel Bueno thought that Jesus Christ did not come to believe that God exists,[47] which is what lead him to ask Ángela to pray not only for his own incredulity, but also for the incredulity of Jesus Christ:

> And then, the last general Communion which our saint was to give! When he came to my brother to give him the Host —his hand steady this time— just after the liturgical "… *in vitam aeternam*", he bent down and whispered to him: "There is no other life but this, no life more eternal… let them dream it eternal… let it be eternal for a few years…". And when he came to me, he said: "Pray, my child, pray for us all". And then, something so extraordinary happened that I carry it now in my heart as the greatest of mysteries: he leant over and said, in a voice which seemed to belong to the other world: "… and pray, too, for our Lord Jesus Christ".[48]

[46] *Saint Manuel Bueno, Martyr*, p. 140 [*San Manuel Bueno, mártir*, po. 1131–1132]. See also *Saint Manuel Bueno, Martyr*, pp. 157–158 [*San Manuel Bueno, mártir*, p. 1141]): "At that moment the fool Blasillo came along our street, crying out his: "My God, my God, why hast Thou forsaken me?" And Lázaro shuddered, as if he had heard the voice of Don Manuel, or even that of Christ".

[47] See *Saint Manuel Bueno, Martyr*, p. 174–175 [*San Manuel Bueno, mártir*, p. 1151]: "Listen, Ángela, once don Manuel told me that there are truths which, though one reveals them to oneself, must be kept from others; and I told him that telling me was the same as telling himself. And then he said, he confessed to me, that he thought that more than one of the great saints, perhaps the very greatest himself, had died without believing in the other life".

[48] *Saint Manuel Bueno, Martyr*, pp. 166–167 [*San Manuel Bueno, mártir*, pp. 1146–1147].

Unamuno went even further than claiming that his conception of religious faith was the one professed by early Christians, affirming that his notion of religious faith was the only one that can make any sense to the common, worldly man: the dogma, the belief in the factual sense regarding theological statements, has no meaning for the concrete man, the "*hombre de carne y hueso*". This point is explicitly made by Unamuno, now using his own voice and not one of his fictional characters', in the short epilogue that accompanies his novel *San Manuel Bueno, mártir*:

> I should like also, since Ángela Carballino introduced her own feelings into the story —I don't know how it could have been otherwise— to comment on her statement to the effect that if Don Manuel and his disciple Lázaro had confessed their convictions to the people, they, the people, would not have understood. Nor, I should like to add, would they have believed the two of them. They would have believed in their works and not in their words. And works stand by themselves, and need no words to back them up. In a village like Valverde de Lucerna one makes one's confession by one's conduct. And as for faith, the people scarcely know what it is, and care less.[49]

VI. Conclusion

Throughout this paper I have argued that the core claims of Unamuno's religious faith are present, in one way or another, in his novel *San Manuel Bueno, mártir.* To sum up, then, it seems reasonable to conclude that the guideline of Unamuno's novel *San Manuel Bueno, mártir* is the expression, in fictional, non-philosophical language, of the conception of religious faith Unamuno had previously defended in his major philosophical work, *Del sentimiento trágico de la vida en los hombres y en los pueblos*.

[49] *Saint Manuel Bueno, martyr*, p. 180 [*San Manuel Bueno, mártir*, pp. 1153–1154].

Schopenhauer's Theory of Agency in Light of His Account of the Affirmation and Negation of the Will

Luís Aguiar de Sousa

Aguiar de Sousa, L. (2019). Schopenhauer's theory of agency in light of his account
of the affirmation and negation of the will.
***Il pensare - rivista di filosofia*, (9), 4-28.**
**http://www.ilpensare.net/il-pensare-n92019-il-mondo-come-volonta-e-
rappresentazione-200-anni-dopo-temi-e-prospettive/**

I.

Although Schopenhauer's explicit theory of agency may seem simple in its outline, when considered more closely it presents, as is usually the case with many topics in Schopenhauer's philosophy, a series of seemingly aporetic contradictions. This is the case not only when we consider his theory of agency apart from the rest of his system, but also, and especially, when we consider it in the broad context of his thinking as a whole. In this paper, I intend to examine this theory in the light of the notions of affirmation and negation of the will.

At first sight, it may seem that the doctrines of affirmation and negation of the will are already part of Schopenhauer's theory of agency, such that when I propose that we establish a relation between them and his theory of action, I'm actually establishing a relation between a part of the doctrine and its whole. It should be noted, however, that Schopenhauer scarcely mentions these doctrines in connection with his discussion of action, character and freedom in §55 of *The World as Will and Representation*[1] and in the *Two Fundamental Problems of Ethics*, just two of the main sources of his account of agency. Accordingly, I propose that we should make a distinction between Schopenhauer's theory of agency in its narrow sense, which includes his accounts of action, character and freedom (which is roughly covered in §55 of WWV I), from his broader account of action (which forms the subject of book IV of WWV I) taken as whole. The former is characterized by the idea of the primacy of the will (or character) over the intellect, of the instrumental character of the latter, and can be seen as an almost naturalistic account of agency. In this context, Schopenhauer also emphasizes the fact that our actions are completely determined by the conjunction of our (individual) character, motives and knowledge of the circumstances in which we find ourselves. His broad theory of agency, on the other hand, which is linked to the doctrines of affirmation, negation and self-knowledge of the will,

[1] Throughout the paper, I will refer to the Hübscher edition of Schopenhauer's works. English translations are taken from the Cambridge Edition of Schopenhauer's works. Page numbers for the Hübscher edition are provided in the margins of the latter.

can be characterized as an "existentialist" one. According to the latter, agency can be traced back to two basic possibilities – the affirmation and the negation of the will, along with their respective degrees – and these are a function of our intuitive and immediate knowledge of the world and its nature. It is also in this context that Schopenhauer puts his doctrine of "transcendental freedom" to use by claiming that the will is ultimately able to "choose" which of the two attitudes it ultimately wants, affirmation or negation.

Accordingly, I will show that if we only take into account Schopenhauer's theory of agency in a narrow sense, Schopenhauer seems to present an irrationalist, determinist and naturalist picture of human agency, whereas if we take into account his doctrines of affirmation and negation of the will, we get a view of agency that allows more space for the role of subjectivity and self-knowledge and that appears to be much closer to "existentialist" accounts of agency, which emphasize authenticity, existential insight, and of course freedom.

II.

The first appearance of Schopenhauer's theory of agency in his published work is in the first edition of *The Fourfold Root of the Principle of Sufficient Reason* (which was Schopenhauer's doctoral dissertation). There are considerable differences between this edition and the better-known third and last edition. In this paper, I will focus on the latter. Here, Schopenhauer considers the will as the fourth type of object of our faculty of cognition, to be distinguished from «*intuitive*, complete, empirical representations»[2], from «space and time» as the «formal part of complete representations»[3], «the forms of the outer and inner senses» (*ibidem*), and «concepts», that is, «abstract representations»[4]. As its own kind of object, the will is subject to a specific form of the principle of sufficient reason. Correlated with this object is also a specific kind

[2] SG, §17, 28.

[3] SG, §35, 130.

[4] SG, §26, 97.

of faculty of the subject: in this case, self-consciousness. The will is, according to Schopenhauer, the object of self-consciousness[5]. As such, the form of the principle of sufficient reason that can be applied to the will is what Schopenhauer calls the «law of motivation», «the principle of sufficient reason of acting»[6]. This entails, for Schopenhauer, that every act of will has its sufficient reason in an antecedent motive. In other words, we feel a priori justified in asking the question «why?» someone does something or other[7]. This is what it means to claim that the principle of sufficient reason applies to the will. In fact, in the *Fourfold Root*, Schopenhauer misstates the view he presents in other passages. As we will see in due time, the principle of sufficient reason finds application only in particular acts of will[8]. According to Schopenhauer, we cannot meaningfully ask why we will in general (*ibidem*). Thus we cannot say that the principle of sufficient reason extends to the "subject of willing" as such.[9] This ambiguity on Schopenhauer's part is also closely related to the ambiguity in which the notion of «object of self-consciousness» is to be found. Although he says that the object of self-consciousness is the subject of willing (*das Subjekt des Wollens*), through self-consciousness or through «inner sense», as he also calls the former, we are only aware of ourselves in time, that is, in the form of succession and not absolutely as we are in ourselves (as a thing-in-itself)[10]. Thus, we could say that the "object" of self-consciousness is not so much the subject of willing as its acts appearing in the form of a succession. This notwithstanding, it must also be added that we are not aware of ourselves as a "pure spectator" would be, indifferent to the fact that, after all, what appears to us in

[5] SG, §41, 140.

[6] SG, §43, 144-5.

[7] SG, §43, 144.

[8] WWV I, §20, 127.

[9] Even the expression "subject of willing" can be contested on Schopenhauer's own grounds, for this expression seems to entail that the subject is somehow detached from his own willing, which as such remains a matter of indifference to him. If we want to be strict, there is no "subject of willing" but at most an "individual character" that expresses herself entirely through her own acts of will.

[10] WWV II, ch. 18, 220-1.

"inner sense" are our own acts of willing. This is why Schopenhauer speaks of an identity between the subject of knowing and the subject of willing in self-consciousness and even calls it the «miracle par excellence»[11]. Through self-consciousness I, as the subject of cognition, am aware of myself as one who wills, that is, as we will see, as an agent who feels responsible for his own acts, who feels these acts as his very own.

Also importantly, that which answers the question "why" regarding acts of will are what Schopenhauer calls "motives". The use of the notion of a "motive", although in many respects similar to our common usage, also deviates from it in other specific respects. Motives can be, for Schopenhauer, either the "real" objects of intuition – Schopenhauer's first class of objects – or abstract concepts and reasonings – Schopenhauer's third class of objects, although even in this latter case they point to real objects and state of affairs. In other words, in its technical sense, Schopenhauer tends to use "motives" for objects rather than, say, a subjective state. To give a concrete example, if we want to employ Schopenhauer's language of motives in this strict sense, we should say that what "causes" or "moves" me to eat is the representation – be it intuitive (perception) or abstract (the thought) – of food or of a particular instance, say a piece of fruit. In this sense, it is not "hunger" that strictly speaking causes me to eat, but rather the sight of food, for example. It is true that, ultimately, I would not feel hungry if it were not for a "will to eat", as part of a more encompassing "will to live", but what explains my particular act of eating is not that I have this "will to eat" but the representation, be it intuitive or abstract, of food.

Thus far, we already know that acts of will have motives and that these are, roughly speaking, "objects". We also know that the principle of sufficient reason for acting cannot be applied to our willing taken as a whole. We cannot legitimately ask why we will rather than not. To go back to our eating example, we cannot ask why we are hungry or are prone to feel hunger, or even why we want life in general. This belongs to what it is to have a willing nature and is without reason or motive; in Schopenhauer's parlance, it is groundless (*grundlos*).

[11] SG, §42, 143.

One of the things that are essential to Schopenhauer's view of agency, and which I have not mentioned until now, is the fact that it entails our embodiment. Besides being cognizing beings, it is only meaningful to think of ourselves as agents as well to the extent that we find ourselves as embodied in the world. This point is already implicit in Schopenhauer's introduction to the second book of WWV. There, Schopenhauer says that it is only possible to investigate the true nature of the world because we are not only the subject of cognition, «a winged cherub's head without a body»[12]. What is implicit in this idea is that beyond being a subject of cognition, for whom the body is only an object, even if an immediate one, we are also the subject of willing. This is in fact one of the points at which Schopenhauer diverts from the view presented in the first edition of the *Fourfold Root*. There, he takes the body to be the "immediate object", both in the sense that it forms the starting point of our perception of the world and in the sense that it is also the starting point of our acting upon the world – it is the first link in the causal chain that begins with the will[13]. Even in his first work, Schopenhauer is quick to point out that we are not really acquainted with the "permanent state" of the will that precedes the causality of the will upon the body as its immediate object[14]. Five years later, in 1919, on the occasion of the publication of WWV, Schopenhauer maintained that the body is the "immediate object of cognition"[15], but he completely transformed his theory on the relation between body and will. As is well known, Schopenhauer claimed from then on that the body and the will are exactly the same thing viewed from two different perspectives. The will is the body seen from the inside, and the body is the will seen from the outside. The *ratio cognoscendi* of this latter claim is the observation[16]

[12] WWV I, §18, 118.

[13] SG1, §45, 74.

[14] SG1, §46, 75.

[15] WWV I, §6, 23-4.

[16] In the *Prize Essay on the Freedom of the Will*, Schopenhauer says that the proposition «I see each act of my will present itself immediately (in a way totally incomprehensible to me) as an action of my body» is «an empirical proposition for the cognizing subject» (FW, II, 22), although in WWV I Schopenhauer says that the identity between body and will is the most immediate cognition, one that cannot really be demonstrated (§18, 122).

that every act of will is at once an act of the body. From this observation, Schopenhauer goes on to show that the "will as a whole" is the same as the "body as a whole" in every respect. At this point, Schopenhauer's conception of the will was broader than our common concept. He claimed that all of our inner states, such as feelings, emotions, etc., fall under the concept of the will. In order to provide proof of this, Schopenhauer points out that «every impression made on my body also instantly and immediately affects my will»; «every violent movement of the will – which is to say affects and passions – agitates the body and disturbs the course of its functioning»[17]; «correspondingly, any effect on the body is instantly and immediately an effect on the will as well»; pleasure and pain are «immediate affections of the will in its appearance, the body»[18]. It is true that by associating the will with the content of our self-consciousness and by broadening its concept to include all kinds of non-cognitive inner feelings, the link between will and action seems to become looser. This link does not need to be abandoned though. Non-cognitive feelings are aimed, in a more or less direct fashion, at corporeal manifestations, and thus at action, even when it does not result in an effective bodily manifestation.[19]

As already indicated, for the law of motivation to be applied to the will, we must presuppose that the latter manifests itself upon motives according to a rule. For Schopenhauer, character is the ultimate presupposition of our motives' eliciting acts of will (or, what is the same, individual actions). Motives are not an absolute explanation of an action. As Schopenhauer puts it, they only explain why the action had to occur at this time and place. Only the fact that the individual will is as it is accounts for the fact that it is liable to act on certain motives and not others. According to Schopenhauer, character is a concept that we form empirically (taking as a starting point our actions or those of any other individual). It lies at the basis of the individual's various actions as the ultimate presupposition of their causal explanation by motives. Although the notion of character is empirical in that we are only get acquainted with anyone's character, including our own, through observation, Schopenhauer thinks that the

[17] WWV I, §20, 128.

[18] WWV I, §18, 120.

[19] C. Janaway makes a similar point, cf. *Schopenhauer on Self and World*, Oxford, Oxford University Press, pp. 221ff.

unity of empirical character is the manifestation of an intelligible character that, as such, is outside of space, time and causality. “Intelligible character” is the a priori unity that is completely inaccessible to us but that we must presuppose as existing outside the forms of appearance (space, time and causality), that is, as a thing-in-itself.[20]

With this background on the notion of character, it is easier to understand Schopenhauer’s point regarding the identity of the act of the will and the “act” of the body. Only the act of the body stamps the act of the will, because only the former is a sure sign of one’s true nature or character.

III.

I will now sketch in more detail Schopenhauer’s theory of agency as it is further developed in paragraph 55 of *The World as Will and Representation*, in the *Essay on the Freedom of the Will*, and in chapter 19 of the second volume of the *World as Will and Representation*.

As I’ve already hinted at, for Schopenhauer, human action ensues from the influence of motives on our character: from the way motives drive our character towards manifestation. The character itself consists of certain permanent drives or aims that, although unbeknownst to us, each of us pursues. Schopenhauer sometimes characterizes character as the innermost rule (or even maxim) of our conduct[21], although it should be borne in mind that this “maxim” is not an abstract principle of our faculty of reason that we consciously choose to follow. For Schopenhauer, human action does not differ in its nature from non-human action. It does not differ even from all “action” in the German sense of *Wirken*. If we take action or acting in the sense of *Wirken* as the genus, human action is but a species of the former concept. What distinguishes human

[20] It should be noted that the qualification “intelligible”, although taken from Kant, is at bottom completely foreign to Schopenhauer’s philosophical intentions. Kant calls it “intelligible” because he considers it to be a “noumenon”, that is, a possible object of non-sensible intuition, a notion that Schopenhauer rejects. For that reason, in the first edition of the *Fourfold Root*, Schopenhauer says it would be preferable to call it “the unintelligible” character (SG1, §46, 76-7).

[21] WWV I, §20, 127; §55, 354.

action from non-human animals' action is the fact that, whereas the latter act on intuitive motives, that is, perceptions, humans are for the most part driven to action by abstract motives, thoughts – that is, concepts and judgements. This latter circumstance gives humans the ability to deliberate, to ponder various motives in an abstract manner, to weigh their influence on our will or character, and to choose accordingly. This gives human action a certain circumspection (*Besonnenheit* in German)[22] that non-human animals lack.

The ability to deliberate, to ponder among various courses of action, is also called the "ability to choose" (*Wahlentscheidung*)[23]. The latter, however, must be carefully distinguished from the empirical "*liberum arbitrium indifferentiae*", that is, a "free choice of indifference", i.e. the idea that in a given situation two opposite actions are possible[24]. This latter kind of freedom is a mere illusion, according to Schopenhauer.

Schopenhauer sees the idea of a "free choice of indifference" as being closely linked to his critique of intellectualism or rationalism: the ancient idea, of Platonic ancestry, that our innermost essence consists in a rational soul[25]. If we were a purely rational or cognitive being – as opposed to what we essentially are, for Schopenhauer: a blind, striving, will – we would will according to our cognition. Life would take the form of a purely intellectual problem. We could become whoever we wanted to be according to what we deemed best[26]. We think we remain "impartial" before the power of motives and weigh them until we reach a completely rational decision. However, to choose among various motives is to see which one is stronger,

[22] The new Cambridge Edition does not translate *Besonnenheit* in a uniform way. It is variously translated as "circumspection", "soundness of mind", "clear-headedness", "thoughtfulness", "mental clarity", "clarity of mind", etc. The problem is that Schopenhauer employs the term mostly in a technical sense. Although all of these meanings may be involved, I think that "circumspection" is the best option because Schopenhauer links *Besonnenheit* with the ability to represent time as a whole, that is, the past and the future alongside the present. Besides this, it also conveys the idea of being prudent or thoughtful in a practical sense, which is also at play here.

[23] WWV I, §55, 351; see also FR, §26, 97.

[24] WWV I, §55, 344.

[25] WWV I, §55, 345.

[26] WWV I, §55, 345.

which one "pulls" us more forcefully. For this reason, there cannot be a purely rational action, that is, one that is driven merely by thinking and weighing reasons for acting (motives). Schopenhauer is thus in complete opposition to Kant's view of agency, in particular Kant's conception of the possibility of a purely rational action, or, in other words, of pure practical reason. In order for action to take place, in order for something to become a motive for us, another side of us, one that is different from the cognitive one, must come into play. Schopenhauer identifies the latter with the will. This dimension of our being cannot be rationally accounted for, and we cannot exert conscious control over it.

Although Schopenhauer does not think our action is conditioned by "blind" causal factors, he thinks that it is necessarily determined by antecedent motives (which are what "cause" actions, according to him). Action ensues with necessity from the solicitation of our individual character by motives. In order for an individual action to be different, we would have to be a different person, have a different nature, in sum have another character. Given one's individual character and the motives that manifest it outwardly, actions cannot be different from what they are. Nevertheless, for Schopenhauer, as for Kant before him, the necessity of actions has validity only at the level of appearances (*Erscheinungen*). Since human beings are as much a part of appearances (*Erscheinung*) as any other natural entities, their actions are just as subject to necessary laws as any other entity in nature. If we consider ourselves as things-in-themselves, however – in other words, if we consider what Kant called our intelligible character, our character as existing outside of time, space and causal relations – then we must consider ourselves free in the sense that our being is not determined or conditioned by anything else. This is a merely negative notion of freedom (this does not mean that Schopenhauer does not have a more positive account of freedom related to his doctrine of the negation of the will, as we will see). Thus, our freedom does not lie where we usually locate it: in the action (which is, on the contrary, thoroughly determined by the motive that elicited it), but rather in our character, in our nature, that is, in our being: «Thus freedom, which cannot be encounterable in the *operari*, must reside in the *esse*»[27]. In WWV, Schopenhauer claims that we have insight into this freedom through

[27] FW, V, 97.

the feeling of the "originality" and "independence" of our acts of will[28]. In FW, where he develops this further, he locates this feeling in our sense of responsibility. The latter concerns what we do, our particular actions, only superficially. In truth, according to Schopenhauer, this feeling is directed at who we are, our innermost nature, that is, our character. This is also related to Schopenhauer's contention that the human being "is his own work"[29] rather than having been made by another (be it God or his parents). In this lies Schopenhauer's deep agreement with existentialist conceptions of the human being and action (like those of Sartre), despite their enormous differences (in particular Schopenhauer's more naturalistic outlook).[30]

Since our being lies in our character, our will, and since the latter (being outside of time as well as all other essential phenomenal forms) does not change, the character or "the will as a whole" is immutable, according to Schopenhauer. Change only occurs in time, and the will is free of time[31]. Schopenhauer interprets the metaphysical immutability of character as the fact that character is inborn (FW, II, 53ff.). Changes in behaviour must be ascribed to changes in our cognition of motives, which for Schopenhauer includes our knowledge of our situation and circumstances. Even if we don't accept this argument because it presupposes Schopenhauer's general metaphysical outlook, if we accept his account of action as the interplay between will or character and the intellect, either we must concede that we have a nature (and then everything that we do must proceed according to this nature) or we must conceive of ourselves as being nothing.[32] For Schopenhauer, if we can be said to exist, we have to have a nature, an

[28] WWV I, §55, 342.

[29] WWV I, §55, 345.

[30] In Schopenhauer, the human being «is his own work prior to any cognition» (WWV I, §55, 345), whereas in Sartre the human being is his own work through being essentially pre-reflectively aware of itself.

[31] WWV I, §55, 344.

[32] The latter idea would later be upheld by Sartre, whose philosophy is thus anticipated by Schopenhauer, at least as a possibility. Since Sartre holds that as the "being-for-itself" we are nothing, as opposed to the "being-in-itself", which is being proper, he also argues, at least in *Being and Nothingness*, that we are radically free. For Schopenhauer, on the other hand, it would be nonsensical to claim that we exist and yet have no nature (which is exactly what Sartre claims, based on his interpretation of Heidegger's idea of the precedence of existence over essence).

essence: «Free will, precisely considered, denotes an existentia without essentia: which means that something would be and at the same time be nothing, which in turn means not be, and is therefore a contradiction»[33].

To this point, I have still failed to mention certain very important aspects of Schopenhauer's doctrine of character. It is not only humans that have a character. Non-human animals also have a character (as does everything else in nature, even non-living beings). The difference is that the character of non-human animals coincides with the character of their respective species (despite Schopenhauer's admission that higher animals show some signs of individuality), whereas humans have, besides the character of the species, an individual character, which is unique to every single person. The fact that humans possess an individual character in addition to a general one entails that, whereas animals immediately exhibit their character or inner nature through action, humans do not. Individuality is also tied to the fact that humans are rational beings[34]. In other words, what they do "in the spur of the moment", unreflectingly, does not adequately express their innermost individual character. Whereas in non-human animals desire tends to pass at once into action, in humans there is a gap between desire and decision.[35] According to Schopenhauer, desire only shows «what *human beings* in general, not the *individual* who experiences this desire, would be able to do»[36]. In other words, desire only manifests the character of the species. Through desire we are drawn into every type of human endeavour. Only those desires that are mediated through the rational deliberative process and issue in a decision are a «sign of individual character»[37]. Even rational decisions can fall short of expressing our innermost individual character, however. If on the one hand reason is a condition of individuality, on the other hand it can also be an impediment to it. The fact that the human being, as a rational being, has

[33] FW, II, 58.

[34] WWV I, §55, 353ff.

[35] M. Koßler, *Schopenhauers Philosophie als Erfahrung des Charakters*, in Dieter Biernbacher-Andreas Lorenz-Leon Miodonski (eds.), *Schopenhauer im Kontext. Deutsch-polnisches Schopenhauer-Symposium 2000*, Würzburg, Königshausen & Neumann 2002, p. 100.

[36] WWV I, §55, 354.

[37] WWV I, §55, 353.

to act according to universal concepts disturbs the manifestation of his individual character[38]. This empirical, innate, character is, to begin with, a «simple drive of nature» (*einfacher Naturtrieb*[39]). As long as we remain unacquainted with our innermost essence, we are doomed to zigzag our way through life[40]. The human predicament is that we do not know our individual characters a priori, and it is a lifelong task to become acquainted with our individual selves, with what each of us basically is as an individual. Our individual character is at first as unknown to us as those of everyone else. We must come to know it through experience. This means that we can be misled about ourselves, about who we really are. For all we know, we may be pursuing certain actions – altruistic ones, for example – only because we believe that a certain reward awaits us in another life. Only those that "acquire character"[41] act in a way that is completely consistent with their individual character. They possess conceptual knowledge of the kind of person they are. They have achieved self-knowledge.

It is not easy to reconcile Schopenhauer's doctrine of "acquired character" with the doctrine that human action always takes place according to the agent's innate character. According to him, character always manifests itself in the agent's life course. The difference is that, whereas those who have acquired character manifest it in a consistent way, the others end up manifesting it by passing through many detours and mistaken paths. Of course, this puts pressure on the idea that actions always reflect our will/character. John Atwell has devised what I think is an ingenuous solution to this problem, however.[42] According to him, actions that are out of character reflect the character of the agent as a specimen of the human race more than his individual character.[43] Here, one could also add that, just as an isolated

[38] WWV I, §27, 181; §55, 357-8.

[39] Cf. WWV I, §55, 357.

[40] *Ibidem*, 358.

[41] WWV I, §55, 357ff.; FW, III, 50.

[42] J. Atwell, *Schopenhauer. The Human Character*, Philadephia, Tempel University Press 1990, pp. 63ff..

[43] *Ibidem*, p. 63: «It follows, I think, that there can be no action "out of character", where that expression means actions explainable without reference to a type of human character; but there can be action "out of character" in that I can do actions that need not be explained by explicit reference to my unique character».

musical note is meaningless apart from the whole set of notes that together make up a certain melody, what is supposed to manifest our character is our life taken as a whole and not an isolated action, which taken by itself is meaningless, or at best ambiguous.[44]

IV.

In this section of the paper, I will examine the notions of affirmation and negation of the will and probe their relation to Schopenhauer's theory of agency as described thus far. The latter is basically what I, at the beginning of this paper, called Schopenhauer's theory of agency in the narrow sense. As already anticipated, in WWV, the notions of the affirmation and negation of the will to life significantly transform the framework of his "narrow" account of agency and shed new light on it.

As indicated in the title of book IV, «with the achievement of self-knowledge, affirmation and negation of the will to life», Schopenhauer introduces the two categories that he will use as a key to understanding the meaning of human action and behaviour: the "affirmation" and the "negation" of the will. According to Schopenhauer, the aim of book IV is precisely to describe the essence of the various modes of behaviour through the guiding thread of these notions in that those modes of behaviour are an expression of the affirmation or negation of the will in their different degrees.

The "affirmation" and the "negation" of the will correspond to what can be called two different and opposite global outlooks on the world and life (even though they are not explicit beliefs). Before we can enter into this issue in more detail, we must still define precisely what affirmation and negation of the will are and why Schopenhauer uses them as clues to interpreting the meaning of the different ways in which humans behave and act.

To affirm the will is simply the same as willing. Since willing is the same as acting, in the sense that it all its manifestations are directly or indirectly connected to action (see section II above), the most simple act

[44] J. Atwell also likens the agent to the "common feature" that belongs to all of his or her actions instead of being a mere bundle of actions (*ibidem*, pp. 38-39)

of will is already an affirmation of the will as such: «the affirmation of the will is the constant willing itself, undisturbed by any cognition, as it fills the lives of human beings in general»[45].

In organic beings, all willing can be fundamentally reduced to the "will" of the individual to preserve its life and the sexual "will", which can be seen as the ultimate goal of the individual, that is, to contribute to preserving its species:

> «The basic theme of all the various acts of will is the satisfaction of needs that are inseparable from the healthy existence of the body, are already expressed in it, and can be reduced to the preservation of the individual and the propagation of the species.»[46]

As Schopenhauer puts it, all willing is will to life. Will to life is a mere "pleonasm"[47]; the two expressions are synonymous. This does not mean that life is the "object" of willing, as if it were its ultimate, conscious, motive. As Schopenhauer makes clear in §29 of the second book of WWV, the will has no aim, end or goal. That the will is will to life means, rather, that what unconsciously propels willing, its "internal mechanism", is its blind tendency to maintain itself in existence. This means, indirectly, that for Schopenhauer the core of our existence lays in our purely biological side. Culture only hides this true nature of ours, and most of the human activities that compose our existence in civilized society – perhaps with the sole exception of artistic contemplation and creation – aim at filling the void generated by the fact that our fundamental will to life is satisfied.[48] The will to life is a drive for the maintenance of life, or simply a drive for existence for its own sake. Of course, in Schopenhauer's grander metaphysical scheme of things, the will to "biological life" is but a particular case of the metaphysical "will" to objectivation or existence,

[45] WWV I, §60, 385.

[46] WWV I, §60, 385.

[47] WWV I, §54, 323-4.

[48] See WWV I, §57, 369 where Schopenhauer says that "boredom" is the root of sociability.

which operates in natural forces as they strive for matter and "compete" with each other to "tak[e] hold" of it[49].

Thus, affirming life in this "natural" sense does not require an explicit stance on our part, as individuals, towards life. The latter is rather something that we pursue for the most part without being aware of it. (It can be anticipated that to "negate the will" can take on the meaning of simply ceasing to will, not willing).[50]

Now, despite the fact that, as we have already seen, humans are distinct from animals inasmuch as their agency is not solely determined by intuitive, present motives and involves an ability to choose among different abstract motives, up to now we may have the impression that the individual will as a whole (or character) is an ineluctable fact: that each has his or her individual character, and there is nothing that can be done about it. As we also saw, however, even if we cannot replace our individual character or transform it in any way, Schopenhauer claims that we are "transcendentally free". We saw that this meant, at first merely negatively, that the will/character is unconditioned. This idea then became more concrete inasmuch as we remain responsible for what we are and a fortiori for what we do (since what we do ensues from, or expresses, what we are). The other way Schopenhauer puts the idea of "transcendental freedom" to use is directly linked to the ideas of affirmation and negation of the will. For Schopenhauer, my acting the way I do entails not only that I am responsible for what I am but also that I "affirm" the will. Furthermore, it is also with reference to transcendental freedom that Schopenhauer introduces the possibility of "negation of the will" and "self-abolition [*Selbstaufhebung*] of character". If there were no "intelligible freedom", it would not be possible to abolish my character, to negate the will.

The claim that the account of character presupposes the affirmation or negation of the will may suggest that action, the natural expression of our individual character, presupposes that we have previously made a "decision" to affirm the will. However, at least in the case of what I called

[49] WWV I, §27, 174-5; §28, 192.

[50] In a passage from his later works, Schopenhauer speaks of the alternative between affirmation and negation of the will as the alternative between *velle* (willing) and *nolle* (not willing). See PP II, §161, 331.

the "natural affirmation of the will", we do not make that decision at any time, and nor does Schopenhauer hold such a view. Nonetheless, it should be noted that this natural "condition" is already one of affirmation of the will. Here, we must go a little bit deeper than Schopenhauer himself and make a subtler distinction within the concept of "affirmation of the will". To affirm the will, as Schopenhauer makes clear in the passage where he introduces the concept for the first time, is also to will life «consciously, deliberately, and with cognition»:

> «The will affirms itself, which means that while in its objectivity (i.e. in the world and life) its own essence is given to it completely and distinctly as representation, this cognition is no impediment to its willing; rather, consciously, deliberately, and with cognition, it wills the life that it thus recognizes as such, just as it did as a blind urge before it had this cognition.»[51]

This attitude differs from the first in that it entails a kind of reiteration of what we already are, a kind of conscious choice of ourselves, of our ultimate nature. What distinguishes it is the fact that the conscious and deliberate affirmation of the will involves some degree of self-knowledge, of self-transparency as will. Although animals possess consciousness, this does not mean that they affirm life in the sense that is at stake here. Schopenhauer makes clear that the will to life (the instinct of self-preservation and the sexual instinct) does not depend on consciousness, much less on reflectively deeming life to be objectively worth living[52].

As for the negation of the will, it has its roots in seeing through the illusion of individuation that veils the will as a "thing-in-itself", that is, in identifying ourselves to a greater or lesser degree, and feeling one, with the will as a whole, which as such is one and the same in every being. Now, it is not just the negation of the will that involves the ability to rise above individuation. The "conscious and deliberate" affirmation

[51] WWV I, §54, 336.

[52] Cf., for example, WWV II, ch. 28, 402.

of the will already entails an overcoming of the individual point of view, an identification with nature as a whole, to the extent that it is a will to life. What is at stake in this latter sense of affirmation of the will is thus the affirmation of our existence as a will to life. The life that was affirmed without consciousness, deliberation, and cognition is from now on not only accepted but wanted as such. From a more substantive point of view, Schopenhauer equates this perspective with what he believes to be the point of view of the Stoics and of Spinoza. (We can also see this as an anticipation of Nietzsche's point of view.) For Schopenhauer, to affirm life is to become conscious of the eternity or immortality of the will to life, to take some comfort in it, that is, to overcome our entrenched fear of death, of ever losing our individual self:

> «Someone who has thoroughly integrated the truths stated so far into his way of thinking, without at the same time having any personal experience or far-reaching insight into the continuous suffering that is essential to all life; someone, rather, who is perfectly happy and content with life and who, after calm reflection, could wish that his life as he has experienced it so far would be of endless duration, or of perpetually new recurrence, and whose thirst for life is so great that he would gladly and willingly take on all the pain and hardships that life is subject to in return for its pleasures; such a person would stand "with firm, strong bones on the well-grounded, enduring earth", and would have nothing to fear: armed with the knowledge that we have given him, he would look at death with indifference as it rushed towards him on the wings of time, regarding it as a false illusion, an impotent phantom, frightening to the weak, but powerless against anyone who knows that he himself is that will whose objectivation or image is the whole world, and to which, for this reason, life and the present will always remain certainties, the true and only form of appearance of the will; the thought of an infinite past or future without him can hold no horror for him, since he regards this as an empty illusion and the web of

> māyā, and thus has as little to fear from death as the sun has to fear from the night.»[53]

Although Schopenhauer does not in any way relate ethics to this affirmation of the will – rather, he relates ethical life to negation of the will and affirmation of the will to evil – since it involves self-knowledge, overcoming individuality and identification with the whole, the way of life of a conscious and deliberate affirmation of the will seems to be "ethically" superior to that of the mere "natural" affirmation of the will. The reason Schopenhauer does not emphasize the "ethical" character of the affirmation of the will seems to be parallel to the reason he does not see any ethical dimension in the acquisition of character (although the latter surely seems to have it). Those who affirm the will with full consciousness still fall short of self-knowledge, that is, of a complete and thorough knowledge of the nature of the will. That is, one of the reasons the negation of the will is ethically superior to the affirmation of the will seems to lie in the fact that the former involves a higher degree of self-knowledge. Although they can comfort themselves with the fact that «for the will to life, life is a certainty, and as long as we are filled with life-will, we do not need to worry about our existence, even in the face of death»[54], those who affirm the will still fall short of the insight that «continuous suffering is essential to all life». In this way, negation of the will seems to be a higher point of view than affirmation of the will solely by the fact that it has a higher cognitive value – those who deny the will have deeper insight into the true nature of the world. In the case of the negation of the will, this insight is, of course, the intuitive grasping of the pessimistic thesis that life is not worth living.

Furthermore, the cognition that is involved in the "conscious and deliberate" affirmation of the will to life and, as we will see, also in the negation of the will, is cognition of essential aspects of the world as will in different degrees, as opposed to the cognition involved in the kind of action that ensues from "blind" affirmation of the will (which is what is described by what I have been calling the "narrow theory of agency"). The former

[53] WWV I, §54, 334-5.

[54] WWV I, §54, 324.

kind of cognition seems, at base, to be identical with that kind of cognition that Schopenhauer introduces at the beginning of book III under the name of "cognition of Platonic Ideas". It must be taken into account that the account of the cognition of Platonic Ideas given in the introduction to the third book[55] is not necessarily specific to aesthetic knowledge and creation and can be seen to be at play in book IV as well, although Schopenhauer never goes into detail on this topic in book IV.[56] He does, however, explicitly remark that cognition of Platonic Ideas is at the basis not only of artistic creation but also of ethical life, and even philosophy: «[b]oth philosophy and art take this cognition [cognition of Ideas] as their point of departure, as does that state of mind which alone leads to true holiness and redemption from the world, as we will discover in this Book»[57].

In acting in accordance with the intuition of Platonic Ideas, in a sense we do not cognize as individuals anymore, but rather from the point of view of the whole, *sub species aeternitatis*, as Spinoza would put it, or as the "pure subject of cognition", as Schopenhauer puts it. This cognition involves overcoming individuation and recognizing ourselves to a certain extent as the same will that is the essence of everything and of the world in general. This is why those who affirm the will consciously, deliberately and with cognition do not fear for themselves as individuals, that is, do not fear death and know that «for the will to life, life is a certainty»[58]. However, contrary to what happens in aesthetic cognition, we do not remain in a purely contemplative attitude. Rather, we act, or, if we come to negate the will, we cease to act, on account of that cognition.

Although Schopenhauer seems to attribute some degree of reflection to the conscious and deliberate affirmation of life, he insists everywhere else that the kind of "cognition" that is at play in affirmation, but especially in the negation of the will, is not the product of reflection, of reason, and

[55] WWV I, §§30-35.

[56] This has already been pointed out by some commentators. See, for example, D. Hamlyn, *Schopenhauer. The Arguments of the Philosophers*, London, Routledge & Kegan Paul, 1980, p. 150 and R. Malter, *Transzendental Philosophie und Metaphysik des Willens*, Stutttgart-Bad Cannstaat, Frommann-Holzboog, 1991, p. 376.

[57] WWV I, §53, 323.

[58] WWV I, §54, 324.

has the character of a global insight into the essence of the world and life. Very much like aesthetic productions, it cannot be properly expressed in words, in abstract concepts[59]. Just as the artistic genius represents his or her vision of the Platonic ideas through the production of a work of art, the one Schopenhauer characterizes as a "genius in the ethical sense"[60] expresses his "vision" through deeds. The parallelism goes so far that Schopenhauer claims that the ethical genius is completely unable to put the vision that guides his conduct into words. For that reason, he resorts to all kinds of fictitious explanations and dogmas in order to account for his action[61]. Perhaps for this reason it is the main task of the philosopher to describe this practical insight in abstract concepts: «our philosophical efforts can extend only to an interpretation and explanation of human action and the innermost essence and content of the very different and even conflicting maxims which are its living expression»[62].

V.

The negation of the will can assume two main forms: that of ethical action proper and that of asceticism. The latter can also be called negation of the will in the strict sense. (This is not the place for a thorough discussion of the different forms of negation of the will and reflection on their identity and differences. Below, I will have the opportunity to briefly discuss the relation between ethics and asceticism.) Ethical action, on its own, can be divided again into acts of justice and acts of altruism (*Menschenliebe*).[63] These, according to Schopenhauer, have their origin

[59] WWV I, §54, 336; §68, 453.

[60] WWV I, §68, 468. The term "ethical genius", as I am using it here, includes not only those who Schopenhauer calls moral or virtuous persons but also all those who are guided by a cognition of the whole, including those who affirm the will "consciously and deliberately".

[61] WWV I, §66, 435, 436.

[62] WWR I, §53, 321.

[63] Because it connotes a mere feeling towards another human being rather than the idea of acting for her sake, I depart from the new Cambridge translation's choice to render *Menschenliebe* as "loving kindness".

not in the use of practical reason or in the state's coercive power, but in an intuition that sees through the principle of individuation, an insight into the fundamental identity of all beings as will. In the *World as Will and Representation*, and more explicitly in *The Prize Essay On the Basis of Morals*, Schopenhauer also identifies the feeling of compassion (*Mitleid*) towards the other as the form of seeing through the principle of individuation that is at play in ethical action.

The affirmation of the will does not remain at the boundaries of our own body, does not limit itself to the activities that are essential to the preservation of one's own life and the satisfaction of the sexual instinct. It naturally tends to overstep the boundaries of the individual body and negate the will/body of the other, for example when I use it in any way to pursue my own ends and interests. Actions thereby cease to be morally neutral.[64] They acquire a moral overtone. The negation of the will/body of the other is the essence of the phenomenon of wrongdoing (*Unrecht*), according to Schopenhauer. For that reason, justice consists in refraining from negating the will/body of the other. Justice has a merely negative status, inasmuch as it is the mere negation of wrongdoing[65]. It should be noted that in order to keep the affirmation of the will within the boundaries of my own body a certain negation of the will is required, and the latter has its roots, as we have already seen, in seeing through the principle of individuation, in intuiting that the separation between me and the other is not absolute[66]. When this intuition goes deeper, we feel compelled to perform acts of altruism, by which we try positively to relieve the other from his pain. For Schopenhauer, altruism – true, unselfish love – is always compassion[67]. This is so because, according to Schopenhauer, only the pain and suffering of the other and not, say, his joy, prompts us to action. When we act compassionately, we see through the principle of individuation and for that reason feel identified with the other, feel his or her pain as our very own, and take pains to relieve and ameliorate it. Here, it is not so much as if we lose all sense of individuality,

[64] J. Atwell, *Schopenhauer. The Human Character*, *op. cit.*, p. 95.

[65] WWV I, §62, 400; §66, 437.

[66] Cf. especially WWV I, §66, 437-8.

[67] *Mitleid*; cf. WWV I, §67, 443-4.

but rather that each of us as individuals will identify with other specific individual wills.[68] Of course, those who choose altruism as a way of life do not identify with this or that particular other, but rather with every possible other, and in this sense can be said to identify with the will as a whole. It could also be asked to what extent altruism is a negation of the will if it does not involve any kind of suspension of action, if on the contrary it involves acting for the sake of another or others. To this question, it might be replied that negation of the will can, to a certain extent, be seen as equivalent to negation of the individual will.[69] Thus, inasmuch as the altruistic will acts for the sake of another will, it must cease to act for the sake of its own interests, its own well-being and woe; in other words, it must negate its own will. Of course, this still leaves much to be explained and answered.[70]

Schopenhauer included altruism under the category of negation of the will in part because he saw it as being on a continuum with asceticism, as if altruism, when radically pursued, led to asceticism. He speaks of the «transition from virtue to asceticism»[71] and says that the source of altruism and asceticism is the same[72]. However, this should not obscure the fact that there are also very clear distinctions between both forms of negation

[68] In the *Prize Essay on the Basis of Morals*, Schopenhauer further clarifies the phenomenology of compassion and its "paradoxical", even "miraculous", nature. According to him, we feel the pain or suffering of the other without losing our sense that it is we who are feeling this and not precisely the other (GM, §16, 211-2).

[69] This can be seen as the answer to the objection that altruism is another form of egoism. For this objection, see J. Young, *Willing and Unwilling: A Study in the Philosophy of Arthur Schopenhauer*, Dordrecht, Martinus Nijhoff Publishers 1987, pp. 115ff. and J. Young, *Schopenhauer*, Abingdon, Routledge, 2005, pp. 182f.

[70] J. Atwell argues that compassionate action contradicts the identity of body and will (*Schopenhauer. The Human Character*, *op. cit.*, pp. 142, 183-4, 208-9). However, as Atwell himself acknowledges, if we construe compassionate actions as a negation of the (individual) will, there is no contradiction. In the latter case, one must view compassionate actions as something that transcends the natural order of things, where egoistic agency prevails (*ibidem*, pp. 98, 100). The question remains, however, what this non-individual will amounts to and whether it can still be called an instance of willing, as Atwell does when he labels it «objective willing» (*ibidem*, pp. 182, 209).

[71] WWV I, §68, 449f.

[72] WWV I, §68, 447.

(perhaps only asceticism can be called a negation of the will in a strict sense). Asceticism is also based on cognition's being able to triumph over the will. This cognition involves not only (as was the case with justice and altruism) seeing through individuation, feeling oneself to be one with the rest of the world, but also grasping in a purely intuitive manner the meaninglessness of the human condition, the futility of all human endeavours. Asceticism is, as it were, the doctrine of pessimism translated into practice.

Schopenhauer acknowledges that besides «mere cognized suffering», there is a second path towards asceticism. The latter often takes place as a consequence of «suffering felt by oneself»[73]. In the end, however, even when he is contemplating this second possibility, Schopenhauer is quick to remark that negation of the will in this case is not a mere "effect" of suffering (in which case it would not be an appearance of freedom), but rather ensues from looking at one's own particular episode of suffering as embodying the true nature of life (that is, as a "Platonic Idea" in Schopenhauer's technical sense):

> «He only becomes truly awe-inspiring when he lifts his gaze from the particular to the universal, when he views his own suffering as a mere example of the whole and, becoming a genius in the ethical sense, treats it as one case in a thousand, so that the whole of life, seen essentially as suffering, brings him to the point of resignation.»[74]

This is an occasion to briefly return to the discussion of freedom. According to Schopenhauer, there is only one instance where freedom manifests itself in appearance. This is the case when "abolition" (*Aufhebung*) of the will takes place[75] as a consequence of the complete negation of the will. When the will abolishes itself, the body still manifests it, for after all it is nothing but objectified will, but the organism no longer finds itself in a state of willing. In this case, freedom manifests itself directly in appearance, according to Schopenhauer. The problem is that this appears

[73] WWV I, §68, 463.

[74] WWV I, §68, 468.

[75] WWV I, §68, 467; §69, 472; §70, 476ff.

to contradict the idea that every appearance is subject to the principle of sufficient reason and, as such, is necessary. In the case of the negation of the will, we appear to have something that lacks sufficient reason. Schopenhauer admits this contradiction outright. He adds, however, that this merely conceptual contradiction mirrors the real one, that of the appearance of a will that no longer wills[76]. Schopenhauer also says that the «key to reconciling these contradictions» lies in the fact that negation of the will involves an «altered mode of cognition»:

> «The key to reconciling these contradictions is that the state in which the character is removed from the power of the motive does not proceed immediately from the will, but rather from an altered mode of cognition. As long as we are only dealing with cognition that is caught up in the principium individuationis and follows the principle of sufficient reason, the motive has an irresistible force; but when we see through the principium individuationis, we immediately recognize the Ideas, indeed the essence of things in themselves, as being in everything the same will, and from this cognition comes a universal tranquillizer of willing; individual motives become ineffective, because the mode of cognition that corresponds to them retreats, obscured by an entirely different mode of cognition.»[77]

This altered mode of cognition corresponds to a cognition of Platonic Ideas, as opposed to cognition of motives. Insofar as they are related to action, the Platonic Ideas are not motives but what Schopenhauer calls a "tranquillizer" (*Quietiv*). Although Schopenhauer does not avoid resorting to the principle of sufficient reason when he suggests that negation occurs as a consequence of our "altered mode of cognition", we could perhaps frame things differently by returning to how he describes the methodological approach pursued in book IV. There, he says that different modes of conduct are an *expression* of a "living cognition":

[76] WWV I, §70, 477.

[77] WWV I, §70, 477.

> «Both [affirmation and negation of the will] take cognition as their point of departure – not an abstract cognition that is expressed verbally, but rather a living cognition that is expressed only through deeds and behaviour and remains independent of dogmas which, as abstract cognition, are preoccupations of reason.»[78]

According to the view Schopenhauer expresses in this passage, the negation of the will does not happen as a *consequence* of a certain cognition but is instead its *expression*. The same goes for the other forms of negation and for all forms of affirmation of the will. Each conduct expresses a certain (metaphysical) view of the world, even if the agent herself is not aware of it in most cases.[79] In this way, Schopenhauer's project in book IV can be envisioned as a hermeneutics of different modes of conduct, as the project of bringing to light the different "cognitions" involved in the fundamental types of behaviour.[80]

To complicate things further, Schopenhauer – not so much in the first volume of WWV but in the second, and also in GM – speaks of

[78] WWV I, §54, 336.

[79] S. Shapshay argues that there is a "Kantian ghost" of intelligible causality hovering over Schopenhauer's work after his 1814 dissertation. Shapshay highlights in particular the role that intellect plays in overcoming the will or character in aesthetic experience, in particular in the experience of the sublime and in asceticism, etc. Shapshay construes this as a remnant of Kant's theory of freedom. What I think Shaphsay overlooks is that this overturning of the will's primacy has nothing to do with the Kantian model of the spontaneous, rational agency of intelligible character. As Schopenhauer makes clear, the cognition that is relevant to the negation of the will is not abstract cognition of reason but rather a type of "practical insight" that is expressed by deeds alone. See S. Shapshay, "Schopenhauer's Early Fourfold Root and the Ghost of Kantian Freedom", in D. V. Auweele, J. Head (eds.), *Schopenhauer's Fourfold Root*, Abingdon, Routledge, 2017, pp. 80-98. For an interpretation that, like Shapshay's, locates the roots of Schopenhauer's theory of freedom and negation of the will in Kant's theory of freedom, see R. Wicks, "Kant's Theory of Freedom in the Fourfold Root as the Progenitor of Schopenhauer's Metaphysics of Will", in D. V. Auweele, J. Head (eds.), *Schopenhauer's Fourfold Root*, Abingdon, Routledge, 2017, pp. 199-212.

[80] According to J. Atwell, every agent has what he calls a "behavioral metaphysics", «in that everyone, in virtue of his or her behavior and moral character is (say) logically committed to some theory of ultimate reality» (*Schopenhauer. The Human Character*, *op. cit.*, p. 116).

an «ethical difference of characters»[81] and claims that each character is determined by a unique mixture of three incentives (*Triebfeder*), each of which is present in a different degree. These incentives are egoism, compassion and malice[82]. In WWV II[83], he speaks of a fourth incentive, that of seeking one's own woe, which he posits as the root of ascetic practices. According to this, affirming one's individual character does not necessarily ensue in egoistic actions. One may have a good (compassionate) character, a good will, and even the negation of the will can be "naturally" explained as an inner tendency of the person in question. Whereas in WWV I the value of morality lies in its "cognitive" value, in the fact that moral, and especially ascetic, conduct expresses deeper insight into the true nature of things, in GM non-egoistic actions are presented as a mere fact of human nature.

It must be admitted that there is no easy way to reconcile Schopenhauer's original presentation in WWV I with that in GM. One could argue along the lines that, since GM's view does not presuppose Schopenhauer's metaphysics and is merely empirical, WWV I must be seen as expressing Schopenhauer's definitive view on the matter. Here, I will only draw attention to the fact that what from one point of view can be traced back to a certain fact – for example, a certain type of character – from another, supposedly deeper, point of view can be seen as the expression of a certain cognition. Through his character and behaviour, for example, the egoist expresses the absolute reality of individuation. The altruist, for his part, expresses the view that individuation is not absolute and comes to see himself in others. In this way, we can trace the notion of good character back to the possession of a certain lived metaphysics. The structure of the *Prize Essay on the Basis of Morals* confirms this interpretation. In that work, Schopenhauer starts by exhibiting the existence of a moral incentive through which our actions aim toward the good of others. Further on, however, Schopenhauer also points to its ultimate condition of possibility, that is, the ultimate identity of all beings and the illusory character of individuation. Thus, the moral

[81] GM, §20, 249.

[82] GM, §20, 252-3.

[83] WWV II, ch. 48, 697, note.

incentive is traced back to the metaphysical insight regarding the unity of everything that exists.

All this notwithstanding, the idea that cognition saves us from our willing condition must be qualified. It is true that cognition is a necessary means of reaching redemption for Schopenhauer, but the will must ultimately be responsible for itself, and thus for its condition in this world. Schopenhauer himself says that «the effect of the tranquillizer is ultimately also an act of the freedom of the will»[84] and that the blame for not being able to see through individuation must ultimately be placed on the will[85]. This problem must ultimately be traced back to the idea that we do not know the will as a thing-in-itself as such, but only its appearance as affirmation of the will. What the will may be besides this remains completely unknown to us[86]. Understanding this, however, would involve a thorough discussion of Schopenhauer's theory of ultimate reality and the metaphysical status of the will.

[84] WWV I, §70, 478-9.

[85] See WWV II, ch. 47, 690.

[86] See PP II, §161, 331.

Abbreviations of Schopenhauer's Works

FW=Über die Freiheit des Willens in *Sämtliche Werke*, vol. 4, *Die zwei Grundprobleme der Ethik*, Wiesbaden, Brockhaus 1972 (*On the Freedom of the Will* in *The Two Fundamental Problems of Ethics*, *The Cambridge Edition of Schopenhauer's Works*, Eng. trans. by C. Janaway, Cambridge, Cambridge University Press 2009.)

GM=Über die Grundlage der Moral in *Sämtliche Werke*, vol. 4, *Die zwei Grundprobleme der Ethik*, Wiesbaden: Brockhaus 1972 (*The Two Fundamental Problems of Ethics*, *The Cambridge Edition of Schopenhauer's Works*, Eng. trans. by C. Janaway, Cambridge, Cambridge University Press 2009.)

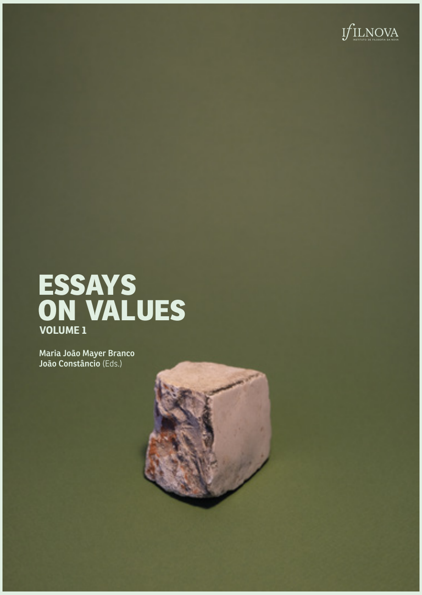
PP II=*Parerga und Paralipomena*, vol. 2, *Sämtliche Werke*, vol. 6, Wiesbaden, Brockhaus 1972 (*Parerga and Paralipomena*, vol. 2, *The Cambridge Edition of Schopenhauer's Works*, Eng. trans. by A. dal Caro-C. Janaway Cambridge, Cambridge University Press 2015.)

SG=Über den vierfachen Wurzel des Satzes vom zureichenden Grund in *Sämtliche Werke*, vol. 1, *Schriften zur Erkenntnistheorie*, Wiesbaden, Brockhaus 1972 (in *On the Fourfold Root of the Principle of Sufficient Reason and Other Writings*, *The Cambridge Edition of Schopenhauer's Works*, Cambridge, Cambridge University Press 2012.)

SG1=Über den vierfachen Wurzel des Satzes vom zureichenden Grund (1813) in *Sämtliche Werke*, vol. 7, *Dissertation. Gestrichenes. Zitate. Register*, Wiesbaden, Brockhaus 1972 (in *On the Fourfold Root of the Principle of Sufficient Reason and Other Writings*, *The Cambridge Edition of Schopenhauer's Works*, Eng. Trans. by D. Cartwright-E. Erdmann-C. Janaway, Cambridge, Cambridge University Press 2012.)

WWV I=*Die Welt als Wille und Vorstellung*, vol. 1, *Sämtliche Werke*, vol. 2, Wiesbaden, Brockhaus 1972 (translated in English as *The World as Will and Representation*, vol. 1, *The Cambridge Edition of Schopenhauer's Works*, Cambridge, Cambridge University Press 2010.)

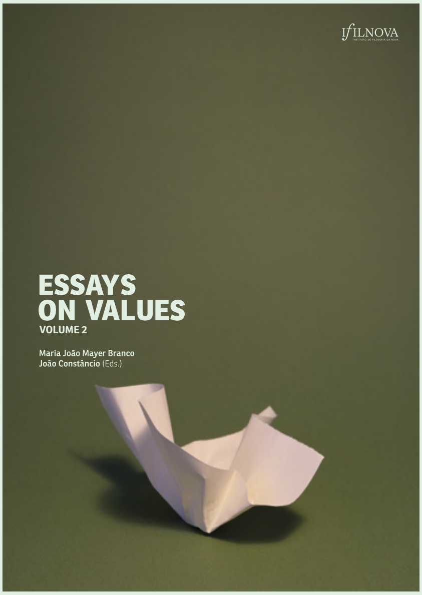
WWV II=*Die Welt als Wille und Vorstellung*, vol. 2, *Sämtliche Werke*, vol. 3, Wiesbaden, Brockhaus 1972 (*The World as Will and Representation*, vol. 2, *The Cambridge Edition of Schopenhauer's Works*, Eng. Trans. by J. Norman-A. Welchman-C. Janaway, Cambridge, Cambridge University Press 2018.)



Volume 1

WHAT ARE VALUES?
NIETZSCHE STUDIES
WITTGENSTEIN STUDIES
ANCIENT PHILOSOPHY

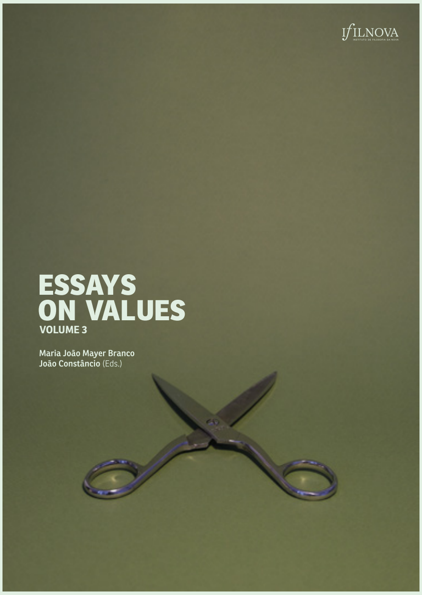
Susana Cadilha and Vítor Guerreiro
João Constâncio
Nuno Fonseca
Alexandra Dias Fortes
Maria Filomena Molder
Eric H. Rast
Maria João Mayer Branco
Marta Faustino
Pietro Gori
Nuno Venturinha
Robert Vinten
Benedetta Zavatta
Paulo Alexandre Lima
Hélder Telo



Volume 2

AESTHETICS
FILM STUDIES

Maile Colbert
Nélio Conceição
Ana Falcato
Ana Godinho
João Lemos
Bartholomew Ryan
Tatiana Salem Levy
Gabriele De Angelis and Emma De Angelis
Stefanie Baumann
Patrícia Castello Branco
Susana Nascimento Duarte
Paolo Stellino
Susana Viegas



Volume 3

ARGUMENTATION AND LANGUAGE
ETHICS AND POLITICAL PHILOSOPHY
EMOTIONS, EMBODIMENT AND AGENCY

Marcin Lewinski and Pedro Abreu
Dima Mohammed and Maria Grazia Rossi
Maria Grazia Rossi
Giulia Terzian and Maria Inés Corbalán
Erik Bordeleau
Filipe Nobre Faria and Sandra Dzenis
Regina Queiroz
Giovanbattista Tusa
António de Castro Caeiro
Robert W. Clowes and Gloria Andrada
Fabrizio Macagno, Chrysi Rapanta,
Elisabeth Mayweg-Paus
and Mercè Garcia-Milà
Dina Mendonça
Alberto Oya
Luís Aguiar de Sousa

ESSAYS ON VALUES 3

ARGUMENTATION AND LANGUAGE
ETHICS AND POLITICAL PHILOSOPHY
EMOTIONS, EMBODIMENT AND AGENCY

IfILNOVA

NOVAFCSH
FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS
UNIVERSIDADE NOVA DE LISBOA

fct
Fundação
para a Ciência
e a Tecnologia