Contents

Research Articles

Book Review

Organon F 29 (1) 2022: 2–25
https://doi.org/10.31577/orgf.2022.29101

ISSN 2585-7150 (online)
ISSN 1335-0668 (print)

RESEARCH ARTICLE

Partial Compatibilism: Free Will in the Light of Moral Experience

David Peroutka*

Received: 11 June 2020 / Revised: 1 December 2020 / Accepted: 25 December 2020

Abstract: Partial compatibilism says that there are basically two kinds of freedom of the will: some free volitions cannot be determined, while others can. My methodological choice is to examine what assumptions will appear necessary if we want to take seriously—and make understandable—our ordinary moral life. Sometimes, typically when we feel guilty about a choice of ours, we are sure enough that we, at the considered moment, actually could have taken a different option. At other times, however, typically when we are aware of some unquestionable moral reasons for a certain choice, we may perceive our choice as voluntary and free in spite of the fact that it is, in the given situation, unthinkable for us to choose otherwise than we actually do (there are situations when responsible agents, because of their strong moral reasons/motives, cannot choose differently). The assumption that experiences of the first kind are not always mistaken excludes our world being deterministic. Yet free will and determinism go together in some of those possible worlds which contain only the second kind of free volitions. Partial compatibilism represents a 'third way' between standard compatibilism and incompatibilism, a way to solve that old dilemma.

* Jan Evangelista Purkyně University
https://orcid.org/0000-0001-9844-3106
Department of Philosophy and Humanities, Faculty of Arts, Jan Evangelista Purkyně University, Pasteurova 13, 400 96 Ústí nad Labem, Czech Republic.
david.peroutka.ocd@seznam.cz

Keywords: Compatibilism; free will; moral experience; volitional necessity.

1. Introduction

When we feel guilty about a decision we have made, we normally suppose that we could have chosen otherwise. Such a kind of free choice seems to involve the existence of alternate possibilities. At other times, however, strong moral reasons drive us to make a wholehearted decision such that, although we perceive that choice as voluntary and free, it is unthinkable for us to choose differently. There seem to be some free choices such that the person in question can (under the same set of conditions) choose otherwise—and there seem to be some others, free as well, such that she cannot. Partial compatibilism, which I am going to propose and explicate, fully recognizes both kinds of situation.

This theory differs from both *incompatibilism* and *standard compatibilism*.[1] Unlike incompatibilism, partial compatibilism admits that there is a possible world where determinism[2] and existence of free will go together. Even in the actual world, arguably, we sometimes make our decision freely despite the fact that we are unable to choose otherwise in the given situation. There seem to be, therefore, some 'compatibilist' possible worlds, namely some of those in which *all* free volitions are of such determined character (I will consider such a world in several paragraphs preceding the conclusion of this paper).

On the other hand, the proposed theory admits, unlike standard compatibilism, that some instances of freedom that can be observed in our *actual* world, namely the cases of moral guilt, are of a kind *not* compatible with determinism. It means that not *every* but only *some* free volitions are

[1] Under the term 'standard compatibilism' I include all compatibilist theories which do not suppose the existence of some acts of volitional freedom incompatible with determinism.

[2] Determinism in the broadest sense of the term says that "everything which happens, happens necessarily; it could not have happened otherwise" (Campbell 1997, 22). Causal or nomological determinism specifies that "every event is necessitated by antecedent events and conditions together with the laws of nature" (Hoefer 2016).

compatible with determinism. That is why I call this theory 'partial'[3] compatibilism.

Before I offer arguments in favour of the claim that some free volitions can occur necessarily, I shall explain my motivation for the thesis that some free volitions cannot be determined and cannot occur necessarily. Both steps, however, need the following preliminary remark on the problem of compatibilism.

The earlier form of standard compatibilism, classic compatibilism (born in the epoch of Enlightenment),[4] often claimed that your action is free if it corresponds to your volition (if you *do* what you *want* to do).[5] This is certainly an understandable account of free *action*. The only trouble is the persisting need to explain what makes *volition* free. Classic compatibilism offered its version of the principle of alternate possibilities.[6] Even if determinism is true, the free agent could have acted otherwise than he actually did. To say that I could have acted otherwise is to claim that I would have acted otherwise *if I had so chosen*.[7] But again, this kind of answer, although

[3] This label relates to the Aristotelian concept of 'partial statement'. "A partial statement (…) asserts that something holds of at least some part of a class, without specifying how large a part it might be…" (Whitaker 1996, 89).

[4] By the term 'classic' compatibilism I generally mean those compatibilist theories which did not reject the principle of alternate possibilities. John Locke, David Hume, George Edward Moore or Alfred Jules Ayer (among others) can be considered protagonists of this kind of approach. Classic compatibilism was the predominant form of compatibilism before Harry Frankfurt's attack on the principle of alternate possibilities (Frankfurt 1969).

[5] "By liberty, then, we can only mean a power of acting or not acting, according to the determinations of the will; that is, if we choose to remain at rest, we may; if we choose to move, we also may" (Hume 1975, VIII, 1, § 23, p. 95). Hume seems to follow John Locke who states: "[T]he Idea of Liberty, is the Idea of a Power in any Agent to do or forbear any particular Action, according to the determination or thought of the mind, whereby either of them is preferr'd to the other (…)" (Locke 1975, II, 21, § 8, p. 237).

[6] The principle of alternate possibilities says that someone is morally responsible for what he has done only if he was able to do otherwise.

[7] Derk Pereboom attributes such an account of freedom to G. E. Moore and A. J. Ayer (Pereboom 2013, 615). Cf. Ayer: "[T]o say that I could have acted otherwise is to say, first, that I should have acted otherwise if I had so chosen; secondly,

explaining how I could have *acted* otherwise, does not explain how I could ever have *chosen* otherwise.

With regard to this distinction I will speak hereinafter of 'choosing' or 'deciding otherwise', rather than of 'acting otherwise'. When freedom of the will is explored, I assess such terminology as more exact.

In the rest of my paper I will firstly explain why partial compatibilism is just *partial*, and, secondly, why it is still a *compatibilism*.

2. What 'ought implies can' implies

Although I believe (with standard compatibilists and some source incompatibilists) that the essence (or definition) of freedom cannot be found in the possibility of choosing otherwise (in the existence of alternate possibilities), I am still inclined to think (unlike standard compatibilists) that some ascriptions of freedom do entail also the attribution of the possibility to choose otherwise.

It has to be noted that I will not present here a fully developed defence of this latter proposition. The existing debate on the topic is extensive and includes (inter alia) the problem of Frakfurt-style counter-examples. Such questions could hardly be answered in brief; and a single paper cannot deal in detail with every question connected with its subject. Consequently, the idea of my paper (as a whole) is developed in a 'hypothetical' or 'conditional' way: the minimal sense is that *even if* freedom of some decision implies the possibility of alternative choice, free will is still compatible with determinism. In what follows, however, I will concretize the main motivation for taking the 'incompatibilist' protasis of this conditional sentence seriously into account.

To be guilty, and, thereupon, also blameworthy[8] for a choice, means that such a choice should not have been made. But whenever we *should*

---

that my action was voluntary in the sense in which the actions, say, of the kleptomaniac are not; and thirdly, that nobody compelled me to choose as I did: and these three conditions may very well be fulfilled. When they are fulfilled, I may be said to have acted freely" (1972, 282). (Cf. Moore 1912, Chapter 6).

[8] It is open to dispute whether the two predicates, "guilty" and "blameworthy", are interchangeable or not. Perhaps one may be blameworthy e.g. for her (innate)

have chosen to act in a certain way, and are guilty because we have not so chosen, we *could* have so chosen. We cannot, for example, fairly blame a person for failing to perform an act the person is, in fact, unable to perform (Mellema 2004, 40). In the case of some decisions—at least those which we assess to be morally wrong—we seem to inevitably suppose that the chooser in question *ought* and (therefore) *could* have chosen otherwise. In the case of morally wrong volitions we might renounce the principle of alternate possibilities[9] only if we were ready to give up the ought-implies-can principle[10] as well (Widerker 1991).

Now the question arises whether or not to believe in the ought-implies-can principle. And I admit that it does not apply universally. Let us remember the distinction, used in late scholasticism, between 'formal sin' and 'material sin'.[11] Formal sin involves guilt whereas mere material sin does not. (Let us hereinafter use the term "guilty" to describe a person who is not only legally, but morally responsible for a morally wrong choice.) Let us assume, for example, that an insane murderer has been—owing to his mental illness—unable to choose otherwise. Then his 'sin' was a 'material' one, but not a 'formal' one. And here we have a counterexample against the ought-implies-can principle. The transgressor in question arguably *should* have chosen otherwise (nobody is allowed to kill innocent and non-attacking people) despite the fact that he *could not* have decided so (owing to his psychological state).[12]

---

bad character without being guilty of it. Our belief that people are blameworthy for their bad traits "does not commit us to holding (…) that people are responsible for [them]" (Sher 2006, 69). Nevertheless, even if it were true that blameworthiness does not always entail guilt, it would be still true that guilt (in the moral sense) entails some blameworthiness.

[9] The principle of alternate possibilities is to be understood in the sense that someone is responsible for his choice only if he could choose otherwise.

[10] The ought-implies-can maxime says that we can be obliged to make only those choices that we are able to make (there can be no obligation to do something one cannot do).

[11] "Peccatum materiale", "peccatum formale" (Cathrein 1915, §98, p. 73).

[12] Cf. the "psychopath case" considered by Julia Driver, i.e., "an example of an agent that is not a moral agent, though is morally appraisable, and the appropriate

In the range of 'formal sins', however, i.e., whenever we surmise that a personal guilt is involved, our assumption that the culprit should have chosen otherwise does imply that he also could have (or else he would not be guilty). In this specific sense I consider the ought-implies-can principle to be an intuitively acceptable axiom of moral theory.[13] And, furthermore, if ought implies can *and* if there are instances in which we 'formally sin' by *not deciding as we ought*, then there are instances in which we can choose otherwise than we actually do. In the actual world, therefore, there are some instances of freedom incompatible with determinism.

It may be further objected that the reason why people are sometimes unable to make an appropriate choice is that they simply lack the motivation needed. In such cases the transgressors, although they cannot take the right option, do usually seem (unlike the above mentioned insane murderer) to be guilty.

An answer can be that if we "know that at some time an agent could not have avoided lacking the motivation required for performing some morally exemplary action", then it would be mistaken to claim that she *ought* to have performed that action at that time (Pereboom 2014, 140). This answer, in my view, is not completely satisfactory. Even if we assume that the agent *at the moment of the considered decision-making* could not have avoided lacking the morally required motivation in question, it does not follow that it would not be possible *tout court* for her to avoid such motivational deficiency: she perhaps could have *in previous times* better formed her motivational moral character, her conscience and will (comp. Kane 2005, 129–131). It seems true, however, that in cases when a transgressor was not able to have done even this (consider e.g. an ill-bred child), or, generally, in cases when she never could *in any relevant sense* have chosen otherwise than she actually did, her fault is but a 'material' one. In other

subject of blame, even though not morally responsible in virtue of lacking the relevant agential capacities…" (Driver 2015, 171).

[13] Perhaps, it might even be argued that the ought implies can principle is as analytically true as, for example, the statements that blue is a colour and children are not adult. Negations of such "analytical" truths just "have no sense" (Grice and Strawson 1956, 150–151).

words, in the realm of 'formal sins' the essential link between 'ought' and 'can' remains untouched even within cases of motivational deficiency.

The ought-implies-can principle is famously threatened by Frankfurt's counter-example, by his story about Jones and Black (with which I assume my reader to be acquainted).[14] And contemporary standard compatibilism, such as John Martin Fischer's semi-compatibilism, heavily relies on this strategy.[15] It is "the basic intuitions elicited by the Frankfurt-type cases which show", in J. M. Fischer's view, "that the most natural justification of the ought-implies-can maxim is faulty" (Fischer 2003, 248).

Frankfurtian reasoning is, however, more disputable than conclusive. Its defenders have been asked, for instance, whether the scenario takes place in a deterministic world or in an indeterministic one. If Jones acts under indeterminism: how could Black ever learn (before Jones's decision is actually made) what Jones is going or not going to decide? If it is, on the contrary, under determinism, then a Frankfurtian compatibilist cannot show (without begging the question against the incompatibilist) whether Jones's freedom or his moral responsibility. (For a useful survey of the debate, see Garnett 2013; see further Kane 1985, 51; Widerker 1995; Ginet 1996; Goetz 2005; Fischer 2010; Palmer 2014; Cohen 2016.) Due to its questionable character, Frankfurtian reasoning does not seem to reliably rule out the ought-implies-can maxim. The question remains open; and it is, therefore, open to us to keep the 'partial' measure of compatibilism.

My partial compatibilism is in some respects similar to Susan Wolf's "asymmetrical" Reason View: Regarding the case in which the agent does just what she ought to do, the Reason View does not require that she have

[14] Let us suppose that Black wants Jones to carry out an action that Jones, morally speaking, certainly should not do. "If it does become clear", Frankfurt says, "that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do." (Black has the power to "manipulate the minute processes of Jones's brain and nervous system".) Black, however, "never has to show his hand because Jones, for reasons of his own, decides to perform and does perform the very action Black wants him to perform" (Frankfurt 1969, 835–836). In such case Jones seems to be guilty despite the fact that he could not have chosen otherwise.

[15] "My motivation for rejecting the 'ought-implies-can' maxim comes from the Frankfurt-type cases" (Fischer 2003, 248).

the ability to do otherwise in order to be free in her choice. But when the agent fails to do what she ought to do, the Reason View does require that she could have done otherwise, namely, that she could have done what she ought (Wolf 1993, 69). "The Reason View is thus committed to the curious claim that being psychologically determined to perform good actions is compatible with deserving praise for them, but that being psychologically determined to perform bad actions is not compatible with deserving blame" (ibid. 79).

There is, however, a difference between Wolf's Reason View and partial compatibilism. A determinism which excludes freedom of wrong actions is, in Wolf's view, *psychological* determinism, i.e., "the thesis that all psychological events are uniquely and wholly determined by a conjunction of laws and states of affairs that are capable of description at the psychological level of explanation" (ibid. 101). "Other forms of determinism", on the other hand, do not contradict the freedom of our choices, not even of our morally wrong choices (ibid. 101–112).[16] By this latter claim Wolf joins standard compatibilism. Unlike me, she believes that free will—also where we assess the volition in question to be morally wrong—is compatible with (global) determinism.[17] I have pointed out, by contrast, that—given the validity of the ought-implies-can principle—our moral responsibility for morally wrong choices requests the kind of freedom which is linked to the principle of alternate possibilities and excluded by whatever impossibility to choose otherwise.

3. Volitional necessity

Having explained in the previous section why partial compatibilism is *partial*, I will give reasons, in the remaining sections of this paper, why

[16] Susan Wolf surprisingly argues that physical determinism goes together with the belief in a real volitional indeterminacy on the psychological level. Her reasoning has been sharply criticised by John Fischer and Mark Ravizza (1992).

[17] "Global determinism is the statement that the world in total is deterministic (however that may be defined) for all times past, present or future. Local determinism is the thesis that determinism does only apply to a certain restricted area, to certain types of processes or at certain times" (Backmann 2013, 11).

partial compatibilism is a *compatibilism*. I will argue that, even if the principle of alternate possibilities sometimes applies, there are still some cases of free volition such that their free character is not connected with any possibility of alternate choice; and, what is more, that there are some 'compatibilist' possible worlds (which are both deterministic and free will containing). The reasoning for this latter claim will constitute an opposition also to those versions of incompatibilism which are ready to renounce the principle of alternate possibilities.

In order to start such considerations, it will be useful to borrow from J. M. Fischer and M. Ravizza the following example of theirs: Matthew, due to his moral conscience, cannot but rescue a drowning child.[18] The example will appear acceptable especially if we expressly add the presupposition that Matthew cannot see *any reason* for not saving the child's life. There is no danger involved for the rescuer (imagine a situation when the best rescue operation consists just in reaching a safety torus at hand, or a swimming ring, down for the child). Nothing prevents or discourages Matthew from the action. He also continues to be a reasonable and conscientious person, not susceptible to panic, etc. Given all those conditions it seems acceptable to assume, with Fischer and Ravizza, that Matthew is not able to choose otherwise.

Although Matthew's choice is undoubtedly driven by significant emotions, such as compassion, desire, and acute worry, the psychological necessity in question does not consist *only* in a force of emotions. If it were so, we could hardly speak of freedom of the will. More likely, it will be useful here to follow on Frankfurt's concept of 'volitional necessity'.

[18] "Here is a case in which an agent is morally responsible for a good action although he could not have done otherwise. Matthew is walking along a beach, looking at the water. He sees a child struggling in the water and he quickly deliberates about the matter, jumps into the water, and rescues the child. We can imagine (…) that if he had considered not trying to save the child, he would have been overwhelmed by literally irresistible guilt feelings which would have caused him to jump into the water and save the child anyway. We simply stipulate that in the alternative sequence the urge to save the child would be genuinely irresistible" (Fischer and Ravizza 1991, 259).

Harry Frankfurt differentiates volitional necessity from the psychological determination which is at work, for example, in the case of an unwilling addict who is forced by his desire to do what he does not want to do. The subject of volitional necessity, by contrast, definitely wants to do just what he is (also emotionally) driven to do. He is unwilling to choose otherwise and this unwillingness "is *itself* something which he is unwilling to alter". Such volitional necessity is not a weakness of the will and is compatible with autonomy and freedom (Frankfurt 1998, 86–88; Frankfurt 1999, 111). Frankfurt connects such volitional necessity mainly with our 'cares' (a care expresses what we love or what is important to us). I think, in addition, that volitional necessity can be linked to our (moral) convictions and judgements.[19] Frankfurt himself, after all, does not separate loves and cares from the broad realm of human rationality; he speaks of a 'volitional rationality': "Violations of volitional rationality (…) are unthinkable" (Frankfurt 2006, 31). It can be said in brief that "for Frankfurt, whereas we ordinarily think of irrationality in terms of transgressing the bounds of what's conceivable (which is delineated by logic), there is also a type of irrationality that amounts to transgressing the bounds of what's *thinkable* (which is delineated by love, or some other volitional necessity)" (Tognattini 2014, 668).

I suggest, in any case, the following general definition of volitional necessity: A person's choice (to act in a certain way) instantiates volitional necessity whenever it is true that even if the person could *act* differently if she had so decided, she would, nonetheless, not be able to *decide* to act differently.[20] My inquiry specifically concerns cases where such volitional necessity is based on the chooser's moral motives. Let us return, in this perspective, to the above presented rescue case.

---

[19] Frankfurt says that "volitional necessity (…) does not derive from a person's moral convictions as such but from the way in which he cares about certain things" (1998, 90). I agree, however, with D. Shoemaker that our moral "self" consists both of our cares and our evaluative judgements (Shoemaker 2015, 115–140).

[20] A person constrained by volitional necessities "may well possess the knowledge and skill required for performing the actions in question; nonetheless, he is unable to perform them. The reason is that he cannot bring himself to do so. It is not that he cannot muster the necessary power. What he cannot muster is the will" (Frankfurt 1999, 111).

Fischer and Ravizza's assumption that Matthew cannot but rescue the child does not seem to contradict their moral appreciation of the case: "Apparently, Matthew is morally responsible—indeed, praiseworthy—for his action, although he could not have done otherwise. Matthew acts freely in saving the child…" (Fischer and Ravizza 1991, 259).

It is thinkable that Matthew chose to rescue the child *not only* under the pressure of emotions but also voluntarily and rationally, i.e., under the guidance of his intelligible value-system. Even if the emotional pressure had been 'overcomeable', he—as an accountable person—could not have forborne to his choice. It was obviously rational to do what he decided to do[21] and *it was evident to him that no other choice was acceptable* in the given situation. In this sense he was not just a 'victim' of his emotions but, despite the determined and necessary character of his choice, he expressed a kind of freedom sufficient for moral responsibility and praiseworthiness.

4. Role of rationality

What kind of freedom, then, is specifically at work in such 'compatibilist' cases? On J. M. Fischer's account of "guidance control" or "freedom"[22] there are "two chief elements": the volition that issues in action must be the "agent's own," and it must be appropriately "reasons-responsive" (Fischer 2007, 78).[23] Let us now consider (more generally) what can be called "rationalist compatibilism", i.e. the view according to which "our freedom is just an expression of our reason" (Pink 2004, 45–46).

Such 'rationalism' is surely open to dispute. Recall Mark Twain's Huckleberry Finn and his friendship with the runaway slave Jim. Huck's moral

[21] "[O]ne explanation for why an agent might not be able to do otherwise is that it is so obviously rational to do what she plans to do and the agent is too rational to ignore that fact" (Wolf 1993, 70).

[22] Fischer seems to identify guidance control with freedom: "[G]uidance control exhausts the 'freedom-relevant' (…) component of moral responsibility. (…) [G]uidance control is all the control (or freedom) necessary for moral responsibility" (Fischer 2006, 107).

[23] The "regulative control", by contrast, "involves access to alternative possibilities (freedom to choose and do otherwise)" (Fischer 2012, 6).

convictions tell him that he should proceed to return the slave to his lawful owner. "Huck believes he is doing wrong in helping Jim escape (…) even though the personal attachment to Jim outweighs the mandate of his conscience" (Bassett 1984, 93). The case became popular among moral philosophers as a counterexample against rationalist accounts of moral responsibility. Huck's choice seems not to be reasons-responsive since it is based on affections and feelings and directly contradicts Huck's normative believes. And yet "there is a strong intuition that Huck is very much praiseworthy for what he does, something that would be impossible if he were not morally responsible for what he does" (Sripada 2016, 1214; Cf. Arpaly and Schroeder 1999).

While I recognize the weight of the objection I still believe that rationality is an essential and defining ingredient of responsibility and freedom of will. If it were not, then the compatibilist should ascribe freedom and moral responsibility also to a dog that voluntarily rescues a child (and, in so doing, perhaps expresses its 'true self':[24] manifests being a good or faithful or friendly dog). Notice, however, that the classic (Aristotelian) concept of voluntariness does not necessarily refer to freedom. A cat chooses to drink milk—rather than tea—quite voluntarily, and yet it lacks freedom: cats are wholly directed by instincts and do not live a moral life.[25] Voluntarily means willingly and intentionally.[26] Although it is true that a dog may rescue a child quite voluntarily, we nonetheless do not take dogs to be free moral agents and do not attribute a *distinctively moral kind* of responsibility to them.[27]

[24] An ethical concept of self-expression has been presented e.g. by Chandra Sripada (2016).

[25] "On Aristotle's telling, animals and children 'share in' voluntary action (EN 1111b8-9), but presumably at least the former do not bear responsibility for their actions" (Klimchuk 2002, 3).

[26] In our context, "intentionality" concerns volitional or affective directedness to a cognized end. Thomas Aquinas says: "It is thus that voluntary action is attributed to irrational animals, in so far as they are moved to an end, through some kind of knowledge." Summa Theologiae I^{a}-IIae, q. 6, a. 2, ad 1.

[27] There is a historical case of a dog worshiped for its (allegedly moral) merits. Such a cult, however, was rather a case of superstition. See Schmitt (1983).

A plausible way for the compatibilist how not to burden animals with moral accountability is to admit that a higher than sensorial knowledge, namely intellective knowledge, essentially concerns that kind of freedom which is needed for moral responsibility. Consider two alternative stories: in one it is Matthew who saves a child in danger whereas in the other one it is a dog. Matthew and the dog may be similar in various respects: both see the child, both feel compassion with her, both cannot but rescue her… Matthew, however, basically differs from the dog in having intellect and knowing (also) intellectively the leading values of his decision.

Although the aim of this paper is not a detailed development of a 'definition' of free will, it will be useful now to outline three defining elements which can be labelled as ownness, voluntariness, and rationality. Firstly, in order for any choice to be considered free it must, from a *psychological* perspective, be the *decider's own*. 'Source compatibilism' wants more: if my choice were a deterministic consequence of the past and the laws of nature, it would not be truly mine (and could never be an instance of free volition).[28] The 'ownness' I am speaking about, by contrast, simply means that I am the chooser in question. Secondly, the choice must be voluntary (unlike, for example, a choice caused by the drug addiction of an "unwilling addict").[29] These two conditions exclude that the *free* decider be a victim of constraint, violence, hypnotic suggestion, and the like. But without adding a further (third) element, the kind of freedom which is linked to moral responsibility should be attributed also to animals and their voluntary choices. So, the moral agent and free decider must, on top of that all, know the leading value of her option not only by the senses, imagination and instincts, but—at least in some measure—also by her intellect.

[28] "The source incompatibilist's position is that this sort of ownership is still not enough. If our motivations are (…) deterministically produced by events to whose occurrence we have not causally contributed—they do not belong to us in the manner required for moral responsibility…" (Shabo 2010, 375).

[29] Voluntariness contradicts not only external constraints but also some internal ones. "If (…) a person is aware of a good reason to do x and still follows his impulse to do y, then he can be said to be impelled by irresistible impulse and hence to act involuntarily. Many kleptomaniacs can be said to act involuntarily…" (Arrington 2001, 121).

When Huck decided to help Jim, he presumably opted for friendship (and perhaps also for some other values: human dignity, solidarity, liberty…). Although Huck lacked the exact value vocabulary, he did not lack some knowledge of his friendship (and other relevant values). And such knowledge was not merely an 'animal' one; Huck's intellect was somehow involved. This can be true independently of what moral premises and conclusions Huck adopted.

The three-membered definition of volitional freedom (as ownness, voluntariness and rationality of volition) is neutral with respect to contingency and determination; it fundamentally permits us to ascribe freedom *both* to contingent and necessary decisions. Your (own voluntary) choice need not be always contingent, i.e., endowed with alternate possibilities, in order to be free; the intellective nature of such a choice is sufficient, according to the 'rationalist' account of freedom, for its being free.

5. Free and yet necessary volitions

In what follows I will offer a more complex argument in favour of the claim that a choice can be psychologically inevitable and yet free.

Let us take for example two persons similar to the poor student Raskolnikov described in Dostoyevsky's *Crime and Punishment*. Let us call them Ivanov and Travkin. Both differ from Raskolnikov by their choice not to kill the avaricious old woman. The first deliberator, Ivanov, makes his decision out of his dilemmatic mental state of incertitude and perplexity. He could have chosen otherwise (was able to murder). His final good decision, due to its contingency, is quite similar to a random result. Conversely Travkin, let us suppose, understands the sense of moral principles so clearly and adheres to them so wholeheartedly and stably that he makes his good decision with necessity.

Since it seems that Travkin's morality surpasses that of Ivanov, my point is that the alleged *universal* validity of the principle of alternate possibilities in some cases divorces, or even puts in conflict, morality and freedom: The more the person, namely Travkin, is virtuous, the less he is free (so it would be, if the necessity of his volition excluded freedom of that volition). And—correspondingly—the less Ivanov is moral, the more he is

free. Indeed, he would be free *in contrast* with (the putatively unfree) Travkin if it were true that freedom *always* needs alternative possibilities. This is a queer rule of proportion. Partial compatibilism avoids such queerness. As can be seen from our example, the principle of alternate possibilities is not universally valid. Consider the range of cases similar to Travkin's choice. From the principle of alternate possibilities it would follow that the higher the ethos of a person who makes a morally obligatory choice is, the lesser her freedom is. Virtuousness deprives us of freedom. This consequence of the principle of alternate possibilities seems to contradict our basic idea of well developed personality in which the morality and the inner freedom constitute a unity and grow together. Partial compatibilism allows for such a harmony.

Now I shall answer two objections that my argument provokes. Firstly, the story of Ivanov and Travkin (unlike the above quoted case of Matthew) seems to be about omissions rather than actions—and the problem of moral responsibility for omissions is not an easy issue.[30] The second objection is Robert Kane's claim that freedom of psychologically inevitable volitions always depends on some preceding undetermined choices.

Regarding the first question, notice that volitions, choices or decisions are not omissions. I have drawn above a distinction between volitions and corresponding actions. It is, however, important to note that volitions, even if they are not actions, are a sort of acts. When Ivanov after a moment of deliberation decided not to perform the action in consideration, he still did perform an act. His final volition itself was an act (namely an act of the will), not an omission; and we are responsible for such acts.

According to the second objection (drawn from Robert Kane's work), the freedom of a decision can be accompanied by an incapacity to decide otherwise exclusively in cases where the psychological necessity of a choice is a consequence of the agent's past "self-forming actions" (Kane 1996, 74) or "self-forming willings" (p. 125), i.e., undetermined will-setting acts in her life-history. Kane describes the "self-forming actions" as "the actions in our lives by which we form our character and motives (i.e., our wills) and make ourselves into the kinds of persons we are" (Kane 2005, 129–131).

[30] See the debates between Frankfurt (1994), Fischer (1997), and Clarke (2014, 119–132).

In this sense Kane treats the Martin Luther case. Luther's decision to pursue his ideas in spite of the ecclesiastical opposition was presumably a token of free will—and yet we should take seriously Luther's statement "Here I stand; I cannot do otherwise".[31] Robert Kane recognizes the possibility of such situation, and yet he insists on the essential connection of freedom with volitional indeterminacy: if Luther's choice were truly free, it must have been preceded by some Luther's *undetermined* "self-forming" or "will-setting" actions.[32]

In Luther's concrete case I am inclined to agree. I do not see, however, why the experience of freedom should not be equally respectable where the psychological necessity in question is a result e.g. of one's natural inclination towards the good or of an innate (and educationally developed) intuition of moral laws,[33] or simply of one's innate (and educationally developed) character, rather than of one's 'libertarian' (undetermined) past self-formation. The Libertarian may argue that our character is *up to us* only to such an extent as it is a result of our own past undetermined self-formation. And moral responsibility must be linked to what is *up to us*. But the Compatibilist may answer back that one's innate moral character is not always an 'excuse from responsibility'. You may, therefore, sometimes express who you are, in a morally relevant sense, regardless of whether or not you have created (or co-created) the character-traits in question by any undetermined self-forming acts.

Interestingly, both conflicting views are well founded on some common moral intuitions. What kind of intuitions, then, should take priority? Perhaps we can observe, with Fischer, that the debates "have issued in what

[31] The words may be genuine (see Bainton 1978, 182).

[32] "All actions done of our own free wills do not have to be undetermined self-forming actions (SFAs) of this kind. (Luther's 'Here I stand' could have been uttered 'of his own free will' even if Luther's will was already settled when he said it.) But if no actions in our lifetimes were of this undetermined self-forming or will-setting kind, then our wills would not be our own free wills and we would not be ultimately responsible for anything we did" (Kane 2005, 130–131).

[33] See the quasi-intuitionist interpretation of Aquinas's natural-law theory developed by John Finnis (2011, 59–99).

some might consider stalemates (…).” We probably cannot “expect knock-down arguments in this realm” (Fischer 2006, 119).

There is, however, a special and yet very common kind of human moral experience which speaks finally in favour of the compatibilist party. There are some voluntary decisions such that *although* the decider was not able, for strong moral reasons, to choose differently, she felt fully free in making her choice or keeping her volition. And *even if* she had learned at that time, or got to believe, that the volitional necessity in question had not been a result of her past ‘libertarian’ self-formation, she still hardly could have been stopped by such a belief from feeling free in her choice. As philosophers, then, we should opt for respecting the way people actually experience freedom of will.

6. A ‘compatibilist’ possible world

Now imagine (or rather think) a world composed only of a kind of pure spirits. Each of them makes only one choice in his lifetime, namely, whether to love and please others for the rest of his life or to hate and harm them. Let us suppose furthermore that, albeit the spirits do deliberate and decide the question, their volitional nature is so constituted that they *necessarily* opt for the first alternative. Moreover, they are not, in any sense, originators of this necessitating nature. Yet their actual option seems, without a doubt, to be morally much better than its opposite, and this is true regardless of whether it is made necessarily or contingently.

According to Thomas Aquinas the “natural necessity” of a volition “does not remove the freedom of will”,[34] because “freedom (…) contradicts the necessity of coercion but not the necessity of natural inclination” of the will.[35] This is the view I defended in the forgoing sections. In the possible world just described, then, *there are* free volitions in spite of the fact that

[34] Thomas Aquinas, Summa Theologiae, I^{a}, q. 82, a. 1, ad 1: “Necessitas autem naturalis non aufert libertatem voluntatis (…).”

[35] Thomas Aquinas, De veritate, q. 22, a. 5, ad s. c. 3: “[L]ibertas (…) opponitur necessitati coactionis, non autem naturalis inclinationis.”

it is a deterministic world. (Indeed, since everything is governed by natural necessity there, a sort of nomological[36] determinism holds in that world.)

Compare the inhabitants of that world with similar creatures in a similar (but non-deterministic) world. These spirits differ from the former ones by being able to take the evil option instead. It perhaps makes sense to say that spirits of the first kind are morally better beings than those of the latter even if we suppose that everybody happens to take the same good decision. In any case, it is sure that spirits of the first kind are morally better compared with a spirit taking the evil option. But importantly, such judgements make sense only if those spirits are endowed with moral responsibility and, consequently, also with the free will which is necessary for being morally responsible. In their world, therefore, we observe both determinism and free will together.

Such a fictional comparison is just a philosophical transcription of a common intuition. Consider two (possible) colleagues who both understand that it is wrong to offend people for no reason. One of them is so good-hearted that, when he deliberates upon the question, he finds himself definitely unable of unjustly offend you. He seems to be a morally better person than the other colleague who, although fortunately refraining from wanton offences, is nonetheless a kind of man able to purposely take such an option. In any case, the first man is (in the relevant respects) morally better compared with a colleague who actually does take that evil option. And again, such moral judgements make sense only if the (determined) volition of the first colleague exemplifies the kind of freedom which is necessary for being morally responsible.

We can sometimes non-mistakenly feel free in making a choice even though this choice is a necessary consequence of our intellective and moral character. And, moreover, if my entire possible world reasoning holds, it is possible, in principle, to be free in such moments *regardless of whether or not we are causally responsible for that character or for its causes*, i.e., whether or not our choice is ultimately determined by something we can control. That is to say, we can non-mistakenly feel free even when the

[36] Laws are not peculiar "Platonic" entities manipulating the behaviour of things; they are descriptions of ways the things regularly function thanks to their dispositions (Cf. Mumford 2004).

freedom relevant requirements of incompatibilism (or source incompatibilism[37]) are not met.

7. Conclusion

Partial compatibilism says that there are basically two kinds of freedom of the will: some free volitions—at least all 'formally sinful' volitions—cannot be determined, while others can. The assumption that experiences of the first kind are not always mistaken probably excludes our world being deterministic (no possible world is both deterministic *and* moral guilt containing). Yet free will and determinism go together in some of those possible worlds which contain only the second kind of free volitions.

Is such possible world discourse just an 'ivory tower' theory divorced from reality? I think it is not. The methodological choice standing behind my inquiry has been to examine what assumptions will appear necessary if we want to take seriously our ordinary moral life. As Charles Taylor (in a different context) says, "What we need to *explain* is people living their lives (…). How can we ever know that humans can be explained by any scientific theory *until* we actually explain how they live their lives in its terms?" (2001, 58).[38] Partial compatibilism is suitable for the ethicist who does not feel attracted to radical revisions of our 'default intuitions' on morality and freedom.

Sometimes, typically when we feel guilty about a choice of ours, we are sure enough that we, at the considered moment, actually could have taken a different option. Standard compatibilism, however, allows that such consciousness may always be false. I have tried to corroborate, by contrast, our

[37] According to source compatibilism your choice is free only if you are the ultimate source or first cause (though not the sole cause) of the choice in question. S. Shabo further explains: "According to source incompatibilists, we can be ultimately responsible for a causally determined decision only if we are ultimately responsible for enough of its causal determinants; responsibility for the former derives from responsibility for the latter…" (Shabo 2010, 358).

[38] Cf. the phenomenological methodology as "hermeneutics of the fundamental phenomena of human life" (Patočka 2016, 127).

intuitive belief that no person really bears *moral* guilt for her choice unless she could have chosen otherwise.

At other times, typically when our moral motives for a choice are strong and unequivocal, we are quite sure that we cannot (*ceteris paribus*) avoid a certain choice; and yet we can in some of those moments experience our choice as voluntary, wholehearted, and free. Incompatibilism cannot explain such a kind of experience without introducing various superfluous assumptions (such as necessity of our past relevant self-formation, or metaphysics of 'ultimate sourcehood'). I argued, contrary to incompatibilism, that our choice sometimes can occur necessarily due to the fact that we are the kind of persons we are; and yet it can, at the same time, be free regardless of whether or not we are originators of that 'determining' nature (or of its causes).

Partial compatibilism, unlike standard compatibilism and incompatibilism, has the advantage, in my view, of neither casting doubt on nor overly conditioning any of the ways we actually experience our use of free will. I am aware that partial compatibilism may look like a compromise seeking or a 'double-faced' theory, as it is located somehow between standard compatibilism and incompatibilism. I tried to show, however, that it is a sufficiently simple, consistent and defensible position; and that its 'doubleness' is useful if we want to make sense of our moral life and be, as theorists, loyal to the variety of human moral experience.

References

Ayer, Alfred Jules. 1972 (1954). "Freedom and Necessity." In *Philosophical Essays*, 271–284. London and Basingstoke: Macmillan. https://doi.org/10.1007/978-1-349-00132-3_12

Arpaly, Nomy, and Timothy Schroeder. 1999. "Praise, blame and the whole self." *Philosophical Studies* 93 (2): 161–188. https://doi.org/10.1023/A:1004222928272

Arrington, Robert L. 2001. "Advertising and Behavior Control." In *Business Ethics—Critical Perspectives on Business and Management*, Vol. II. Edited by Alan R. Malachowski, 112–125. London and New York: Routledge. https://doi.org/10.1007/BF00382800

Backmann, Marius. 2013. *Humean Libertarianism: Outline of a Revisionist Account of the Joint Problem of Free Will, Determinism & Laws of Nature.* Frankfurt: Ontos Verlag. https://doi.org/10.1515/9783110320701

Bainton, Roland H. 1978. *Here I stand: A Life of Martin Luther*. Nashville: Abingdon Press. https://doi.org/10.1177/004057365200800413

Bassett, John Earl. 1984. "Huckleberry Finn: The End Lies in the Beginning." *American Literary Realism, 1870-1910*, 17 (1): 89–98.

Campbell, Robert. 1997. "Philosophy and the Accident." In *Accidents in History: Injuries, Fatalities and Social Relations*. Edited by Roger Cooter and Bill Luckin, 17–34. Amsterdam: Rodopi. https://doi.org/10.1163/9789004418516_005

Cathrein, Victor. 1915. *Philosophia Moralis*. Freiburg im Breisgau: Herder.

Clarke, Randolph. 2014. *Omissions: Agency, Metaphysics, and Responsibility*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199347520.001.0001

Cohen, Yishai. 2016. "Fischer's Deterministic Frankfurt-Style Argument." *Erkenntnis* 82 (1): 121–140. https://doi.org/10.1007/s10670-016-9809-7

Driver, Julia. 2015. "Appraisability, Attributability, and Moral Agency." In *The Nature of Moral Responsibility: New Essays*. Edited by Randolph Clarke, Michael McKenna, Angela M. Smith, 157–174. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199998074.003.0008

Finnis, John. 2011. *Natural Law and Natural Rights*. New York: Oxford University Press.

Fischer, John Martin, and Mark Ravizza. 1991. "Responsibility and Inevitability." *Ethics* 101 (2): 258–278. https://doi.org/10.1086/293288

Fischer, John Martin, and Mark Ravizza. 1992. "Responsibility, Freedom, and Reason" (a review of Wolf's book Freedom within Reason). *Ethics* 102 (2): 386–389. https://doi.org/10.1086/293402

Fischer, John Martin. 1997. "Responsibility, Control, and Omissions." *The Journal of Ethics* 1 (1): 45–64. https://doi.org/10.1023/A:1009707919608

Fischer, John Martin. 2003. "'Ought-Implies-Can', Causal Determinism and Moral Responsibility." *Analysis* 63 (3): 244–250. https://doi.org/10.1093/analys/63.3.244

Fischer, John Martin. 2006. "Responsibility and Self-Expression." In *My Way. Essays on Moral Responsibility*, 106–123. New York: Oxford University Press.

Fischer, John Martin. 2007. "Compatibilism." In: *Four Views on Free Will*. Edited by John Martin Fischer, Robert Kane, Derk Pereboom, Manuel Vargas, 44–84. Malden, Oxford, Carlton: Blackwell Publishing. https://doi.org/10.1111/j.1467-9973.2009.01564.x

Fischer, John Martin. 2010. "The Frankfurt Cases: The Moral of the Stories." *Philosophical Review* 119 (3): 315–336. https://doi.org/10.1215/00318108-2010-002

Fischer, John Martin. 2012. "Deep Control: The Middle Way." In *Deep Control. Essays on Free Will and Value*, 3–29. New York: Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199742981.003.0001

Frankfurt, Harry G. 1969. "Alternate Possibilities and Moral Responsibility." *The Journal of Philosophy* 66 (23): 829–839. https://doi.org/10.2307/2023833

Frankfurt, Harry G. 1994. "An Alleged Asymmetry between Actions and Omissions." *Ethics* 104 (3): 620–623. https://doi.org/10.1086/293633

Frankfurt, Harry G. 1998. *The Importance of What We Care About*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511818172

Frankfurt, Harry G. 1999. "On the Necessity of Ideals." In *Necessity, Volition, and Love*, 108–116. Cambridge, Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9780511624643.010

Frankfurt, Harry G. 2006. *Taking Ourselves Seriously and Getting It Right*. Edited by Debra Satz. Stanford: Stanford University Press.

Garnett, Michael. 2013. "Fischer-style Compatibilism." *Analysis* 73 (2): 387–397. 397, https://doi.org/10.1093/analys/ant018

Ginet, Carl. 1996. "In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Arguments Convincing." *Philosophical Perspectives* 10, 403–17. https://doi.org/10.2307/2216254

Grice, H. Paul, and Peter F. Strawson. 1956. "In Defence of a Dogma." *The Philosophical Review* 65 (2): 141–158. https://doi.org/10.1093/acprof:oso/9780199587292.003.0002

Goetz, Stewart 2005. "Frankfurt-Style Counterexamples and Begging the Question." *Midwest Studies in Philosophy* 29 (1): 83–105. https://doi.org/10.1111/j.1475-4975.2005.00107.x

Hoefer, Carl. 2016. "Causal Determinism" (Spring 2016 Edition). *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta. https://plato.stanford.edu/archives/spr2016/entries/ determinism-causal/ [accessed January 9, 2020].

Hume, David. 1975. *Enquiries concerning Human Understanding and concerning the Principles of Morals*. Edited by Lewis Amherst Selby-Bigge, 3rd edition revised by Peter H. Nidditch. Oxford: Clarendon Press. https://doi.org/10.1093/actrade/9780198245353.book.1

Kane, Robert. 1985. *Free Will and Values*. Albany: State University of New York Press.

Kane, Robert. 1996. *The Significance of Free Will*. New York: Oxford University Press. https://doi.org/10.1093/0195126564.001.0001

Kane, Robert. 2005. *A Contemporary Introduction to Free Will*. New York: Oxford University Press. https://doi.org/10.1093/0195126564.001.0001

Klimchuk, Dennis. 2002. "Aristotle on Necessity and Voluntariness." *History of Philosophy Quarterly* 19 (1): 1–19.

Locke, John. 1975. *An Essay Concerning Human Understanding*. Edited by Peter H. Nidditch. Oxford: Oxford University Press. https://doi.org/10.1093/actrade/9780198243861.book.1

Mellema, Gregory F. 2004. *The Expectations of Morality*. Amsterdam: Rodopi.

Moore, George Edward. 1912. *Ethics*. London: Williams & Norgate.

Mumford, Stephen. 2004. *Laws in Nature*. London: Routledge. https://doi.org/10.4324/9780203458426

Palmer, David. 2014. "Deterministic Frankfurt cases." *Synthese* 191 (16): 3847–3864. https://doi.org/10.1007/s10670-016-9809-7

Patočka, Jan. 2016. *The natural world as a philosophical problem*. Evanston: Northwestern University Press.

Pereboom, Derk. 2013. Free Will. In *The Oxford Handbook of the History of Ethics*. Edited by Roger Crisp, 609–637. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199545971.013.0029

Pereboom, Derk. 2014. *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199685516.001.0001

Pink, Thomas. 2004. *Free Will: A Very Short Introduction*. New York: Oxford University Press. https://doi.org/10.1093/actrade/9780192853585.001.0001

Schmitt, Jean-Claude. 1983. *The Holy Greyhound: Guinefort, Healer of Children Since the Thirteenth Century*. London and New York: Cambridge University Press. https://doi.org/10.1086/ahr/89.4.1063-a

Shabo, Seth. 2010. "Uncompromising Source Incompatibilism." *Philosophy and Phenomenological Research* 80 (2): 349–383. https://doi.org/10.1111/j.1933-1592.2010.00328.x

Sher, George. 2006. *In Praise of Blame*. New York: Oxford University Press. https://doi.org/10.1093/0195187423.001.0001

Shoemaker, David. 2015. "Ecumenical Attributability." In *The Nature of Moral Responsibility*. Edited by Randolph Clarke, Michael McKenna, Angela M. Smith, 115–140. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199998074.003.0006

Sripada, Chandra. 2016. "Self-expression: a deep self theory of moral responsibility." *Philosophical Studies* 173 (5): 1203–1232. https://doi.org/10.1007/s11098-015-0527-9

Taylor, Charles. 2001 (1989). *Sources of the Self. The Making of the Modern Identity*. Cambridge, Massachusetts: Harvard University Press.

Tognazzini, Neal A. 2014. "Reactive Attitudes and Volitional Necessity." *Value Inquiry* 48 (4): 677–689. https://doi.org/10.1007/s10790-014-9464-7

Whitaker, C. W. A. 1996. *Aristotle's De Interpretatione: Contradiction and Dialectic*. Oxford: The Clarendon Press. https://doi.org/10.1093/0199254192.001.0001

Widerker, David. 1991. "Frankfurt on 'Ought Implies Can' and Alternative Possibilities." *Analysis* 51 (4): 222–224. https://doi.org/10.1093/analys/51.4.222

Widerker, David. 1995. "Libertarianism and Frankfurt's Attack on the Principle of Alternative Possibilities." *The Philosophical Review* 104 (2): 247–61. https://doi.org/10.2307/2185979

Wolf, Susan. 1993. *Freedom within Reason*. New York, Oxford: Oxford University Press.

Organon F 29 (1) 2022: 26–46
https://doi.org/10.31577/orgf.2022.29102
ISSN 2585-7150 (online)
ISSN 1335-0668 (print)

RESEARCH ARTICLE

Frege on Identity and Co-Reference

Eros Corazza*

Received: 1 September 2020 / Revised: 26 October 2020 / Accepted: 28 December 2020

Abstract: In "Über Sinn und Bedeutung" (1892) Frege raises a problem concerning identity statements of the form a=b and he criticizes the view he holds in the *Begriffsschrift* (1879, § 8). In building on a suggestion by Perry (2001/12, ch. 7) I will show how Frege's *Begriffsschrift* account can be rescued and how Frege's 1892 criticism of his Begriffsschrift's position somewhat miss the point. Furthermore, the Begriffsschrift's view can be developed to account in quite an elegant way to the so-called Frege's Puzzle without committing to the sense/reference (Sinn/Bedeutung) distinction Frege introduces in "Über Sinn und Bedeutung". To do so we have, though, to give up the idea that all the relevant information conveyed by the utterance of a simple sentence is encapsulated into a single content. I will show of this can be done in adopting a Perry-style pluri-propositionalist model of communication.

Keywords: Frege; Perry, identity co-reference.

1. Introduction

In "Über Sinn und Bedeutung" (1892) Frege raises a problem concerning identity statements of the form $a=b$ and he criticizes the view he holds in

* The University of the Basque Country. Ikerbasque, Basque Foundation for Science. Carleton University
https://orcid.org/0000-0001-8074-7777
The University of the Basque Country UPV-EHU, Donostia, Spain.
eros.corazza@carleton.ca

the *Begriffsschrift* (1879, § 8). In building on a suggestion by Perry (2001/12, ch. 7) I will show how Frege's *Begriffsschrift* account can be rescued and how Frege's 1892 criticism of his *Begriffsschrift*'s position somewhat misses the point. Furthermore, the *Begriffsschrift*'s view can be developed to account in quite an elegant way to the so-called Frege's Puzzle without committing to the sense/reference (*Sinn/Bedeutung*) distinction Frege introduces in "Über Sinn und Bedeutung". To do so we have, though, to give up the idea that all the relevant information conveyed by the utterance of a simple sentence is encapsulated into a single content. I will show how this can be done in adopting a Perry-style pluri-propositionalist model of communication.

2. The *Begriffsschrift's* account

The famous passage in the *Begriffsschrift* where Frege discusses identity, reads as follows:

> Equality of content differs from conditionality and negation by relating to names, not to contents. Elsewhere, sign are mere proxies for their contents, and thus any phrase they occur in just expresses a relation between their various contents; but names at once appear *in propria persona* so soon as they are joined together by the symbol of equality of content; for this signifies the circumstance of two names' having the same content. Thus, along with the introduction of a symbol for equality of content, all symbols are necessarily given a double meaning—the same symbols stand now for their own content, now for themselves. (Frege 1879, § 8)[1]

Frege then goes on to introduce the new notation: the three horizontal strokes symbol, '≡', that stands for the identity of content. We need a sign for the identity of content because, in many cases, the same content can but be given by different names insofar as there must be different modes of

[1] I am adopting Geach's translation (in Geach & Black eds. 1952).

determination (*Bestimmmungsweise*) for the same content.[2] To do so he proposes a geometrical example and he concludes that:

> The name B thus has in this case the same content as the name A: and yet we could not antecedently use just one name, for only the answer to the question justify our doing so. The same point is determined in a double way: (i) it is directly given in experience, (ii) it is given as the point B corresponding to the straight line's being perpendicular to the diameter. To each of these two ways of determining it there answers a separate name. The need of a symbol for equality of content thus rests on the following fact: The same content can be fully determined in different ways; and *that*, in a particular case, *the same* content actually is given by *two ways of determining it*, is the content of a *judgment*. Before this judgment is made, we must supply, corresponding to the two ways of determination, two different names for the thing thus determined. The judgment needs to be expressed by means of a symbol for equality of content, joining the two names together. It is clear from this that different names for the same content are not always just a trivial matter of formulation; if they go along with different ways of determining the content, they are relevant to the essential nature of the case. In these circumstances the judgment as to equality of content is, in Kant's sense, synthetic. (Frege 1879, § 8)

In short, modes of determination are what triggers Frege to introduce the notion of identity of content, expressed by '≡'. For, if Frege were simply focusing on the content, as he does when using matemathical examples (see e.g. § 1 of the *Begriffsschrift*), he would merely use the '=' sign.[3] It is because the same content sometimes can be given only *via* different modes of determination, and thus by using two different names, that Frege appeals to the identity of content symbol '≡'.

---

[2] An interesting question would be to investigate how the notion of modes of determination of the *Begriffsschrift* relates to, and somewhat anticipates, the notion of modes of presentation of "Über Sinn und Bedeutung". On this particular question see Simon (1995).

[3] For a discussion on why Frege does not use mathematical examples when discussing the identity of content (expressed by '≡') see May (2001).

What does a statement of the form $a \equiv b$ mean? If we stick with what Frege textually says in the passage quoted: "this signifies the circumstance of two names' having the same content". Thus, when the names '*a*' and '*b*' flank the '≡' sign, it comes to mean that '*a*' and '*b*' have the same content.

In the *Begriffsschrift*, previous to the introduction of the sense/reference (*Sinn/Bedeutung*) distinction in his 1890 essays, Frege assumes that the content of a name is exhausted by what the name stands for. Thus, 'Tully', in the utterance of a simple sentence like "Tully is Roman" stands for Tully, the referent (or object) designated by the tokened name. Given that Tully is Cicero the names 'Tully' and 'Cicero' stand for the same object. There is a difference, though, between an utterance of "Tully is Roman" and one of "Cicero is Roman". It is in order to capture this difference that Frege goes on to introduce the '≡' symbol for identity of content. Let us consider:

(1) Tully is Roman
(2) Cicero is Roman

Since 'Tully' and 'Cicero' have the same content, (1) and (2), express the same content. If we adopt the notion of proposition, we could thus say that (1) and (2) express the same proposition, i.e. that Tully/Cicero is Roman.[4] A constituent of such a proposition is the object (the referent or content of the names appearing in subject position). Consider now a simple, *modus ponens*, inference like:

(3) a. If Tully is Roman, then Tully is European
b. Tully is Roman
c. Therefore: Tully is European

From (3a) and (3b), though, we cannot infer:

(4) a. Therefore: Cicero is European

[4] To be precise, though, we should talk about state-of-affairs or circumstances (as something that can obtain) when talking about the content of an utterance when interpreting Frege's *Begriffsschrift* (see Mendelsohn 1982: 286). For simplicity sake I will talk about propositions, for the main point I am trying to articulate is independent of this particular interpretation of the *Begriffsschrift*.

Yet, 'Tully' and 'Cicero' have the same content. I surmise that Frege was driven by problems like this then he introduced in his logical notation the three stroke sign, '≡'. Thus, to infer (4a) from (3a) and (3b), we have to add the premise that 'Tully' and 'Cicero' have the same content, i.e.:

(5) d. Tully ≡ Cicero[5]

Why did Frege introduce the '≡' symbol and not employ the traditional equality sign, '='? In other words, what is the difference between '≡' and '='? After all, in the first paragraph of the *Begriffsschrift* Frege uses the '=' sign: "This indeterminateness makes it possible to express by means of letters the general validity of propositions; e.g.: (a + b)c = ac + bc".[6] We are thus entitled to assume that in the *Begriffsschrift* Frege operates with two distinct signs, '=' and '≡'.[7] In Frege's *Begriffsschrift* we thus have a

[5] In the last sentence of § 8 of the *Begriffsschrift* Frege claims that ⊢(A≡B) means: "the symbol A and the symbol B have the same conceptual content, so that A can always be replaced by B and conversely". (Frege's notation '⊢' means, roughly, 'it is a fact that'). I ponder that by "conceptual content" Frege means the inferential power names and other expressions exhibit in inferential reasoning and how they can or cannot be substituted salva veritate in such reasoning.

[6] For a detailed discussion of Frege's distinction between '=' and '≡' see Perry (2020). See also Mendelsohn (1982) and May (2001). Perry proposes an interpretation of the *Begriffsschrift* without taking into consideration § 8. He argues that the *Begriffsschrift's* account can be developed to take into considerations the problems Frege pointed toward when he introduced the sense/reference distinction without appealing to Frege's '≡'. My aim is more modest insofar as I think that '=' and '≡' can subsist together in a coherent picture that deals with some of Frege's various insights. Actually, if I am right, both '=' and '≡' must enter the picture if our aim is to develop an account sensitive to the problems Frege's pointed out in "Über Sinn und Bedeutung". For an interesting discussion on how the *Begriffsschrift* relates to Frege's mature work see also Simon (1995) who argues: "in 'On Sense and Reference' the different ways in which a referent is given allow identities to be informative. If we apply the same way of counting levels to *Begriffsschrift* as to 'On Sense and Reference', we indeed found three, not two. We have the sign, its content, and the way in which the content is determined by the sign … If we do not actually have sense in Begriffsschrift, we seem to have the next best thing" (Simon 1995: 133).

[7] Although in the *Begriffsschrift* Frege uses the '=' sign only once (in the first quoted paragraph), in his "Applications of the 'Conceptual Notation'" (written on

difference between (i) identity (expressed by '=') and (ii) identity of content (expressed by '≡'). While the former is a relation between things, the latter is a relation between signs. Actually, identity of content cannot hold between objects other than certain linguistic objects. While it makes sense to say that Tully is identical to himself, it does not make sense to say that Tully (the object) has an identity of content to himself, or that he entertains a content identity to himself. While '=' expresses a metaphysical (or ontological) relation, '≡' expresses a linguistic relation. I reckon that Frege's "Über Sinn und Bedeutung" criticism of his § 8 of the *Begriffsschrift* somewhat blurs this distinction. To this criticism I now turn.

3. The "Über Sinn und Bedeutung"'s interpretation

The famous controversial passage under discussion starts as follows:

> Equality gives rise to challenging questions which are not altogether easy to answer. Is it a relation? A relation between objects, or between names or signs of objects? In my *Begriffsschrift* I assumed the latter. (Frege 1892, 56-7)[8]

In this passage Frege considers only the '=' sign and suggests that in his *Begriffsschrift* he understood it as a relation between signs. Yet, as we saw, in the paragraph under discussion of the *Begriffsschrift*, Frege did not discuss '='; rather, he used the three stroke sign, '≡'. I am not accusing Frege of misunderstanding between linguistic (grammatical) phenomena and metaphysical ones. All I am claiming is that Frege somewhat "misunderstood" himself, i.e., that the interpretation offered in "Über Sinn und Bedeutung" of § 8 of the *Begriffsschrift* is misleading.[9] To be precise, Frege seems to

the same year) in his notations Frege utilizes both '=' and '≡' (see Frege 1879b, 204-8). For instance, on page 205, Frege writes "we can regard 'u+1 = v' as a function of u and v …" and in a single notation on the same page Frege uses both '='and '≡'. This is further evidence that at the time of the *Begriffsschrift* Frege operated with both signs.

[8] I am adopting Black's translation (in Geach & Black eds. 1952).

[9] To my knowledge the first who pointed out that the Frege of "Über Sinn und Bedeutung" mischaracterizes the view he holds in the *Begriffsschrift* is Angelleli who

argue that in the *Begriffsschrift* he interpreted what is ordinarily expressed as identity statements strictly in terms of his content-identity symbol and, thus, he adopted a particular analysis of statements involving '='. The question we now face is why Frege, and most of his scholars following him, thought that Frege's "Über Sinn und Bedeutung" view of § 8 of the *Begriffsschrift's* is the correct one?[10] My guess is that this misunderstanding is based on the fact that the Frege of "Über Sinn und Bedeutung", unlike the Frege of the *Begriffsschrift*, (like most of his followers) thought that one ought to operate with either '=' or '≡' and that the two signs cannot coexist when we come to explain the problems Frege was after.

In "Über Sinn und Bedeutung" Frege presents us with the following problem when he rebuts his *Begriffsschrift*'s account. I quote the whole paragraph:

> Equality gives rise to challenging questions which are not altogether easy to answer. Is it a relation? A relation between objects, or between names or signs of objects? In my *Begriffsschrift* I assumed the latter. The reason which seems to favour this are the following: $a=a$ and $a=b$ are obviously statements of different

suggests that in his later work Frege is not faithfully reproducing the semantics view he holds in the Begriffsschrift: "Frege himself in this respect has done injustice to his own text of 1879" (Angelleli 1967, 40). In his criticism of the Begriffsschrift's in this passage, Frege undermines (or dismisses) the notion of modes of determination, the very notion that triggered him to introduce the identity of content symbol, '≡'.

[10] The standard interpretation of Frege's first paragraph of "Über Sinn und Bedeutung" has recently been questioned. Thau & Caplan (2001), for instance, argue that Frege never dismissed his Begriffsschrift's interpretation of identity statements. For a criticism of Thau & Caplan's interpretation see Dickie (2008) who argues that Frege's *Begriffsschrift* solution differs from the one proposed in "Über Sinn und Bedeutung" insofar as Frege was concerned with two distinct puzzles. While in the *Begriffsschrift* Frege focuses on why a rational agent can understand two co-referential terms without realizing that they co-refer, in "Über Sinn und Bedeutung" Frege is concerned with inferences, i.e. why in a deductive proof we can provide justification in moving from the premises to the conclusion based on self-evident logical reasoning. The replacement of a term with a co-referential (or co-extensive) one in such a logical deduction may make the proof not logically self-evident and, thus, the two terms differs in cognitive value.

> cognitive value; $a=a$ hold a priori … while statements of the form $a=b$ often contain very valuable extension of knowledge and cannot always be established a priori. … Now if we were to regard equality as a relation between that which the names 'a' and 'b' designate, it would seem that $a=b$ could not differ from $a=a$ (i.e. provided that $a=b$ is true). A relation would thereby be expressed of a thing to itself, and indeed one in which each thing stands to itself but to no other thing. What we apparently want to state by $a=b$ is that the signs or names 'a' and 'b' designate the same thing, so that those signs themselves would be under discussion; a relation between them would be asserted. But this relation would hold between the names or signs only in so far as they named or designated something. It would be mediated by the connection of each of the two signs with the same designated thing. But this is arbitrary. Nobody can be forbidden to use any arbitrary producible event of object as a sign for something. In that case the sentence $a=b$ would no longer be concerned with the subject matter, but only its mode of designation; we would express no proper knowledge by its means. But in many cases this is just what we want to do. If the sign 'a' is distinguished from the sign 'b' only as an object (here, by means of its shape), not as a sign (i.e. not by the manner in which it designates something), the cognitive value of $a=a$ becomes essentially equal to that of $a=b$, provided $a=b$ is true. (Frege 1892, 56-7)

If we break down this paragraph we have two main notions at work: cognitive significance and identity. Identity, we are told, is a relation "in which each thing stands to itself but to no other thing". In that case, though, we cannot distinguish between statements of the form $a=a$ and statements of the form $a=b$. For, if the latter is a true statement, it would express the very same thing, i.e. that a (or b) is identical to itself. Frege argues that in his *Begriffsschrift* he assumed that in such cases what we assert is a relation between signs or names. But this cannot be the case, for we lose the subject matter and would express no proper knowledge. In uttering "Tully is Cicero" one is not talking about the names 'Tully' and 'Cicero', but about Tully/Cicero. Yet, as we saw, in the famous § 8 of the *Begriffsschrift* Frege does not discusses '=', but '≡'.

I now try to suggest how the two accounts can be combined in dealing with the difference between statements of the form $a=a$ (e.g.: "Tully is Tully") and statements of the form $a=b$ (e.g.: "Tully is Cicero").[11] While in "Über Sinn und Bedeutung" they would be represented as:

(5) Tully = Tully
(6) Tully = Cicero

in the *Begriffsschrift* they would be represented as:

(7) Tully ≡ Tully
(8) Tully ≡ Cicero

(8) reads as: 'Tully' and 'Cicero' have the same content. The two names are, therefore, co-referential. In linguistics we usually express co-referentiality using co-indexation. Hence, (8) can be represented as:

(9) $\text{Tully}_1 = \text{Cicero}_1$

The three stroke sign of the *Begriffsschrift* can thus be represented by the subscript signifying co-referentiality and the latter differs from identity: 'Tully' and 'Cicero' are different names after all. (9) can thus be understood as encompassing both '=' and '≡'. If my understanding is right, then with an utterance of "Tully is Cicero" a speaker/writer conveys two pieces of information: (i) that Tully is identical with Cicero *and* (ii) that 'Tully' and 'Cicero' are co-referential (they have the same content). In so doing we do not loose the subject matter, for we are talking about the object, Tully/Cicero, the subject matter of the utterance; we are talking *of* an object carrying two names.[12] At the same time, though, we also suggest that the names 'Tully' and 'Cicero' have the same content, *viz.* that they co-refer (as it is stressed by them sharing the same subscript). In short, with an utterance of an identity statement of the form $a=b$ we convey two pieces of information.[13]

[11] Given that Frege also consider definite descriptions to be proper names the same story could be told using "The Morning Star is the Evening Star".

[12] For a discussion about identity in "Über Sinn und Bedeutung" and the notion of subject matter, see Corazza & Korta (2015).

[13] For an interesting discussion about the difference between identity and co-reference see May (2012). For a discussion of the difference between the identity of

The obvious question that now comes to mind is: why did Frege not analyze a statement of the form $a=b$ the way I did above? I suspect that Frege could not envisage an interpretation along these lines because he was presupposing that all the relevant information ought to be encompassed into a single content. In the *Begriffsschrift* the utterances "Tully is Tully" and "Tully is Cicero" express the same content, i.e. that the object Tully/Cicero is identical to itself. This is the problem that Frege recognizes in "Über Sinn und Bedeutung". For, if they express the same content, we cannot explain how the first is trivial, while the second may helps us to expand our knowledge. This is the well-known and discussed Frege's puzzle. It is also well-known that to solve this problem, i.e. the difference in cognitive significance between the two utterances, Frege introduces the sense/reference distinction. Though 'Tully' and 'Cicero' have the same reference (stands for the same thing) they express different senses. Senses are the constituents of the thought expressed by an utterance. While in the *Begriffsschrift* "Tully is Tully" and "Tully is Cicero" express the same proposition (have the same content), in "Über Sinn und Bedeutung" they express different thoughts and the latter is the bearer of cognitive significance.

The problem of what is the sense expressed by a tokened name has been largely discussed. It is not my intent to engage in this rich and often controversial debate. My aim is more limited. I merely want to show how we may reconcile our insight from the *Begriffsschrift*'s and "Über Sinn und Bedeutung".[14] To do so, though, we must give up the view that a single utterance comes equipped with a single content, be it a proposition or a

content of the *Begriffsschrift* and the notion of identity Frege develops in his mature period, see May (2001).

[14] As far as I know, the first who suggested that Frege's *Begriffsschrift* account can be made consonant with the one he proposed in "Über Sinn und Bedeutung" is Perry (2001/12, see in particular ch.7, section 3; see also Corazza's 2003 review of Perry's 2001), when he spelled out the critical referentialism framework and hints at how the reflexive content of an utterance captures the Fregean account in the *Begriffsschrift*, while the referential content deals with the problem of the subject matter Frege insists upon in "Über Sinn und Bedeutung". More on this in the next section.

thought. The prevailing view is that an utterance can be (semantically) associated only to one content, be it a thought ("Über Sinn und Bedeutung") or a proposition (*Begriffsschrift*). In the next section I will show how the *Begriffsschrift*'s view can be developed to deal with the problems Frege raises in the first paragraph of "Über Sinn und Bedeutung". To summarize, we can agree with both the *Begriffsschrift*'s position that statements of the form $a=a$ and $a=b$ express the same proposition (have the same content) and the view proposed in "Über Sinn und Bedeutung" that they differ in cognitive significance (express different thoughts). To do so, though, we have to assume that statements like these come equipped with more than a single content or proposition. This can be done in adopting a Perry-style pluri-propositionalist model of communication.

4. Back to the *Begriffsschrift*

To understand Frege's "Über Sinn und Bedeutung" discussion of his account of identity in the *Begriffsschrift* and the way he distinguishes between '≡' and '=' we must take a detour. I suggested that the '≡' symbol represents a linguistic relation, while the '=' sign a metaphysical one. Actually, Frege often answers semantics/grammatical concerns in relying on ontological (or metaphysical) distinctions. At the same time, though, Frege drives ontological distinctions based on the grammatical ones. When it comes to discuss the role of a name (*Eigenname*), for instance, Frege characterizes it as *what designates* an object, while he characterizes an object as *what is designated* by a name. The same holds with predicates or concept-words (*Begriffsswort*). A predicate is what denotes a concept and a concept is what is referred to by a predicate. As Dummett puts it:

> Frege's use of the ontological term 'object' is strictly correlative to his use of the linguistic term 'proper name': whatever a proper name stands for is an object, and to speak of something as an object is to say there is, or at least could be, a proper name which stands for it. The question therefore naturally arises in which realm, the linguistic or the ontological, the principle of classification is to be applied. (Dummett 1973/1981, 55-6)

I conjecture that this is also what happened when Frege discussed identity in the *Begriffsschrift* and in "Über Sinn und Bedeutung". While in the former he focuses on the semantics/grammatical relation, in the latter he focuses on the ontological one. And it is from his ontological perspective that in "Über Sinn und Bedeutung" Frege understands and undermines the identity account he proposes in § 8 of the *Begriffsschrift*. If I am right, though, both accounts can subsist. We have, though, to give up the view that all the relevant information conveyed by the utterance of a sentence is encompassed into a single content, be it a proposition or a thought. Frege, like many of his followers, committed what Barwise & Perry characterized as the fallacy of misplaced information, i.e.: "The idea that all the information in an utterance must come from its interpretation [the proposition expressed] we call the *fallacy of misplaced information*" (Barwise & Perry 1983, 38).

For '≡' and '=' to coexist we must avoid the fallacy of misplaced information. One way to do so is to accept the (Perry-inspired) view that an utterance comes equipped with different contents or truth-conditions.[15] Let me illustrate the framework I endorse that allows us to avoid this fallacy. The position I propose can be characterized as *pluri-propositionalism*. For, a single utterance comes equipped with variegated contents or propositions. This, though, does not amount to say that in producing an utterance a speaker ends up expressing (or saying) a multitude of propositions. It simply means that many propositions (or truth-conditions) are *available* when we come to analyze a communicative interaction. Propositions are abstract entities that, although they have no causal power, play important classificatory roles. This framework can be viewed as a reaction to *mono-propositionalism*, or to what Korta (2007) characterizes as the dogma of mono-propositionalism.[16] That is, the view that, if we discount implicatures and

[15] See Perry's (1988, 2001/12) critical referentialism (see also Korta & Perry's critical pragmatics, 2011).

[16] If my interpretation is correct Mendelsohn's critique of Frege's *Begriffsschrift* theory, i.e. that "names in BG [Begriffsschrift] were systematically ambiguous; they stood for their objects they customarily denoted everywhere save when they occurred at either end of the ec [equality of content, '≡'] symbol, at which place they stood for themselves" (Mendelsohn 1982, 285), does not affect the reconstruction I am

presuppositions, there is one and only one proposition associated with the utterance of a sentence. This proposition is required to play variegated roles such as: representing the semantic content of the utterance, what the speaker said, the proposition expressed, the information transmitted, the content of attitudes (what is referred to by that-clauses), the output of semantics, the input for Gricean reasoning, and so on and so forth. No unique proposition can play all these different roles. In what follows I will offer a brief justification for this conjecture.

Pluri-propositionalism, as I take it, is a hybrid between the *Begriffsschrift* and "Über Sinn und Bedeutung"—i.e. the view that the content of a tokened name is the object it refers to *and* the view that names contribute in *conveying* some descriptive information. Thus, the utterance of a simple sentence containing a proper name, on top of expressing a proposition having the referent of the name as a constituent, also carries information about the way the speaker and/or hearer apprehends this proposition. This descriptive information captures, I will show, what in the *Begriffsschrift* Frege characterizes as the modes of determination of the content and, thus that Frege's *Begriffsschrift* account already has all the relevant tools to deal with the main problems Frege highlights in the first paragraph of "Über Sinn und Bedeutung".

To quickly illustrate the pluri-propositionalist framework I defend, let us consider a simple scenario. When seeing John, Sue tells him: "Your shoe is untied". John thinks "My shoe is untied", and stoops to tie it. A case of observation, leading to communication, leading to action. But what is communicated? The traditional answer is: a proposition, i.e. that John's shoe is untied. But the duties that fall upon this proposition are weighty. It must get at what Sue observed and said, what John understood and thought, and the reason for John's action. Why did not Sue tell John: "John's shoe is untied"? She would have said the same thing after all, *viz.*, that John's shoe is untied. If, instead of addressing John using the possessive 'your', Sue

proposing. For, in whichever utterance names appear they stand for their customary content (object); yet, at the same time, as we will now see, they get mentioned in the reflexive content. The Perry-inspired view I am defending does not assume that names, or utterances for that matter, are ambiguous insofar as they are associated with different contents. Perry's critical referentialism is not an ambiguity thesis.

addressed him using his name, John could well not stop to tie his shoe. For, on top of the fact that in a face-to-face situation (of this sort) it is unconventional to address someone using his or her proper name, John may think that Sue was not telling him that his shoe is untied, but that the shoe of someone else sharing his name is untied. John could also be amnesiac and not know that his name is 'John', and so on and so forth. In processing Sue's utterance "Your shoe is untied" John, as a competent speaker of English, understands that his own shoe is untied. If, instead of talking to John, Sue were talking to Jane, to pass the same message, she could not say "Your shoe is untied", for she would be telling Jane that her shoe is untied. Rather, she would say: "John's shoe is untied". Jane could thus direct her attention toward John without bothering about her own shoe. How can we explain these simple communicative situations that trigger different actions?

The traditional answer is that what we express and grasp in a communicative interchange is a proposition. The search for a single proposition is misguided. There is a structure of related propositions, that are not intrinsically equivalent, but equivalent in the circumstance, that does the job. What Sue sees can be captured by an existential or "Fregean" proposition: There is a man I see and a shoe he wears, and it is untied. But to get at the common element between what she says, "Your shoe is untied", and what John understands, "My shoe is untied", we seem to need a proposition about John, a so-called "Russellian" or "singular" proposition that is not a description of John, but John himself that is the common element. Similarly with Sue telling Jane: "John's shoe is untied". If Jane does not know whom Sue intends to talk about, by being a competent speaker of English and recognizing that 'John' is a proper name, she would nonetheless grasp an existential or "Fregean" proposition: There is someone named 'John' whose shoe is untied. To know whom Sue is talking about, Jane has to identify John and, thus, grasp a singular or "Russellian" proposition that is not a description of John but a proposition with John himself as a constituent.[17]

[17] The idea that a single utterance may express more than one proposition is not new. When distinguishing between tone and sense, Frege already hinted at that: "But whilst the word 'dog' is neutral as between having pleasant or unpleasant associations, the word 'cur' certainly has unpleasant rather than pleasant associations and put us rather in mind of a dog with somewhat unkempt appearances. Even

To summarize, the pluri-propositionalist model can be spelled out, roughly, as follows. Utterances of simple sentences like:

(10) Your shoe is untied
(11) John's shoe is untied

come equipped with various contents. Their analysis starts by distinguishing between the reflexive and the referential (or official) contents. Thus, while (10a) and (11a) constitute the reflexive contents, (10b) and (11b) are the referential contents:

(10) a. There is an individual x the speaker of (10) addresses by uttering the possessive 'your' & the speaker of (10) says that x's shoe is untied
b. That John's shoe is untied

if it is grossly unfair to the dog to think of it in this way, we cannot say that this makes the second sentence false. True, anyone who utters this sentence speaks pejoratively, but this is not part of the thought expressed. What distinguishes the second sentence from the first is of the nature of an interjection. It might be thought that the second sentence does nevertheless tell us more than the first, namely that the speaker has a poor opinion of the dog. In that case, the world 'cur' would contain an entire thought" (Frege 1897, 240-1, italics added). Bach (1999) and Neale (1999), for instance, argued that in uttering a sentence a speaker may say two things at once. Bach argues that in uttering "Tom is rich but he is honest" one expresses two propositions, i.e. (i) that Tom is rich and (ii) that there is a contrast between being rich and being honest. Corazza (2002) argues that utterances containing complex names also express more than one proposition. E.g. "The Virgin Mary is Jesus' mother" expresses the propositions that Mary was Jesus' mother and that Mary was a virgin. This helps us to deal with anaphoric pronouns linked with expressions composing the complex name such as 'she', 'one' and 'that color' in "Little [Red1 Riding Hood2]3 was so-called because she3 wore one2 of that color1". Without denying that in uttering a single sentence a speaker can express more than one proposition, i.e. she can say more than one thing at a time. The pluri-propositionalism I defend following Perry, though, is of a different nature. For it is committed to the view that each utterance comes equipped with various contents and that some of the latter (the reflexive contents) do not pertain to the Gricean what is said or Kaplanian content.

(11) a. There is an individual x and a convention C such that: C is exploited by the speaker of (11); C permits one to designate x with 'John' & the speaker of (11) said that x's shoe is untied
b. That John's shoe is untied

By simply hearing an utterance of (10) or (11), a competent speaker would understand something like (10a) and (11a) even if she is unable to grasp who the speaker is and whom she or he designates with his or her use of 'your' and 'John'. These are the *reflexive* contents of utterances of sentences like (10) and (11). They represent the conditions the referent must fulfill to be the individual the speaker refers to and intends to talk about. What the speaker (in our example, Sue) says, though, is not something about these contents. What she says is something about John's shoe and what she says is true just in case John's shoe is untied. What Sue expresses is the proposition that John's shoe is untied. That is, in uttering (10) or (11) Sue expresses the proposition (10b)/(11b). Since the latter is the same, in uttering either (10) or (11) Sue said the same thing. But she said it in different ways, i.e. in exploiting different conditions that John's shoe, the referent and propositional constituent, must fulfill, in the context of the utterance and communicative exchange, to enter the proposition expressed by Sue.

The traditional philosophical understanding of the truth-conditions of a given declarative utterance are the incremental conditions needed to judge whether it is true or false, once all the linguistic and contextual factors are fixed. In short, in our analysis we start from the product, *viz.* the utterance of a given sentence abstracted away from the context of the utterance. That is, we start from the meaning the utterance inherits from the sentence, the type. In so doing we quantify over meanings. We then proceed to fill in the missing ingredients from the actual circumstances in which the utterance occurs. In our analysis we can see that an utterance conveys many other relevant information. In other words, it is by starting to fill in more and more contextual information that the incremental truth-conditions (the official content) gets computed.[18] This does not mean, though, that a

[18] "It is fair to call these truth-conditions of [the note], because they are conditions such that, were they satisfied, [the note] would be true … they are reflexive conditions, conditions on [the note] itself. The truth-conditions on which philosophers

speaker/hearer ought to be consciously aware of all the processing going from the pure reflexive content to the incremental one. Yet, they play an important *classificatory* role. In particular, they help us to classify what goes on in the speaker/hearer mind when she processes an utterance. In so doing it helps us to deal with problems pertaining to *cognitive* significance.[19]

The utterance of a sentence like (1), "Tully is Roman", can be analyzed as follows:

(12) In uttering 'Tully' one refers to Tully
(13) Tully satisfies 'is Roman'

We can thus cash out the reflexive content of (1) as follows:

(14) There is an individual x and a convention C such that:
  (i) C is exploited by (1)
  (ii) C permits one to designate x with 'Tully'
  (iii) x is Roman

The referential (official) content would correspond to the proposition expressed (roughly, the intuitive what is said or Kaplanian content):

traditionally focus are incremental; they are conditions on the subject matter; that is, what the world beyond the utterance must be like, for the utterance to be true; or, as I like to put it, what else, has to be true, given the linguistic and contextual facts about the utterance … the conditions will not say much about the world independently of [the note]. However the familiar philosophical concept of truth-conditions corresponds to the case in which one knows a lot about [the note], so the incremental, what else must be the case for [the note] to be true, are conditions that pertain to the world outside [the note], not [the note] itself … as you figure out more about [the note], fixing more of its linguistic properties, the conditions that had to be fulfilled for its truth become more focused on the world." (Perry 2001/12, 93-4)

[19] In the hands of the theoretician propositions, qua abstract entities, play an important classificatory role. It is in this sense that reflexive contents help to deal with problems pertaining to cognitive significance, i.e. what is going on in the speaker/hearer mind during a communicative interaction and, thus, what Frege in "Über Sinn und Bedeutung" comes to characterize as the modes of presentations. For more on this see Corazza (2018).

(15) That Tully is Roman

In short, the reflexive content captures Frege's *Begriffsschrift* view that, associated to an expression there must be a mode of determination of its content. In grasping the reflexive content, the hearer can start processing relevant information that may ultimately, if all goes well, enable her to grasp the official or referential content. In a nutshell, as communication goes, we can focus on the variegated contents an utterance can convey. If we now consider Frege's identity statements of the form $a = b$ like:

(16) Tully is Cicero

it can be analyzed as follows:

(17) (i) There is an individual x and an individual y and conventions C and C^* such that:
(ia) C and C^* are exploited by (17)
(ib) C permits one to designate x with 'Tully' while C^* permits one to designate y with 'Cicero'
(ii) $x = y$

(17) represents the reflexive content of (16). In this content the names get mentioned and it is stated that they are co-referential (i.e. have the same content), as the "$x = y$" stresses. Once again, (17) encapsulates the *Begriffsschrift*'s identity of content sign, '≡'. The official or referential content of (16) would simply be that Tully/Cicero is identical to itself. Since these contents, *qua* abstract entities, help us to classify what goes on in one mind they can give a way to deal with Frege's puzzle about cognitive significance and to explain people actions. For this reason, in "Über Sinn und Bedeutung" Frege introduces the notion of sense. Roughly, he assimilates senses to the modes of presentation of the objects referred to. As I take it, reflexive contents are what allows us to classify (from a theoretical viewpoint) how speakers cognize the referents. They help to classify the mental contents cognizers entertain when uttering or hearing a sentence. In that sense, *qua* classifiers of what goes on in speaker/hearer mental realm they help us to deal with problems pertaining to *cognitive* significance. As Kaplan puts it: "We use the manner of presentation, the character, to individuate psychological states, in explaining and predicting action" (Kaplan

1989, 532). It is in that sense that I argued that the reflexive contents are what help us to deal with problems pertaining to the *cognitive* significance of an utterance.

By now it should be clear how the *Begriffsschrift*'s account can be understood to counter the criticism Frege proposes in "Über Sinn und Bedeutung". The *Begriffschrift*'s identity of content sign, '≡', is explained at the lever of the reflexive content, where the names flanking it get mentioned. On the other hand, the identity sign, '=', of "Über Sinn und Bedeutung" gets analyzed at the level of the official or referential content. We can thus combine both accounts without rejecting the *Begriffschrift*'s view. In particular, we can accommodate the view that: "along with the introduction of a symbol for equality of content, all symbols are necessarily given a double meaning—the same symbols stand now for their own content, now for themselves" (*Begriffsschrift*: § 8).

5. Conclusion

I hope to have shown that: (i) in "Über Sinn und Bedeutung" Frege somewhat mischaracterized the view he proposed in the famous § 8 of the *Begriffsschrift*, (ii) The *Begriffsschrift*'s account does not crumble under the criticism Frege proposes in "Über Sinn und Bedeutung", (iii) '=' and "≡" should both enter a plausible picture about communication, (iv) an identity statement like "Tully is Cicero" must be analyzed in appealing to both '=' and '≡'. That is, by adopting what I characterized as the pluri-propositionalist model. It is by dismissing mono-propositionalism that Frege's *Begriffsschrift* account of identity of content and the one he presents in "Über Sinn und Bedeutung", can both be incorporated to deal with a plausible theory of communication and handle some of the problems Frege was after without having to subscribe to the sense/reference distinction.

Acknowledgment

For discussions and/or comments I am grateful to Christopher Genovesi, Kepa Korta, María De Ponte, John Perry and a referee of this journal. Research for this paper has been partly sponsored by the Spanish ministry of economy

and competitivity (FFI2015-63719-P (MINECO/FEDER, UE)); the Spanish ministry of science and innovation (PID2019-106078GB-I00 (MCI/AEI/FEDER, UE)) and the Basque Government (IT1032-16).

References

Angelleli, Ignacio. 1967. *Studies on Gottlob Frege and Traditional Philosophy*. Dordrecht: D. Reidel. ISBN 13: 978-94-017-3175-1

Bach, Kent. 1999. "The Myth of Conventional Implicature". Linguistics and Philosophy 22: 327–66. DOI: 10.1023/A:1005466020243

Barwise, Jon & Perry, John. 1983. *Situations and Attitudes*. Cambridge Mass: MIT Press. ISBN: 0262021897, 9780262021890

Corazza, Eros. 2002. "Description-Names". *Journal of Philosophical Logic* 31 (4): 313–25. DOI: 10.1023/A:1019950905478

Corazza, Eros. 2003. "Review of John Perry 2001. Reference and Reflexivity". Palo Alto: CSLI Publications. *Mind* 112 (445): 171–75. DOI: 10.1093/mind/112.445.171

Corazza, Eros. 2018. "Identity, Doxastic Co-Indexation, and Frege's Puzzle". *Intercultural Pragmatics* 15 (2): 271–90. https://doi.org/10.1093/mind/112.445.171

Corazza, Eros & Korta, Kepa. 2015. "Frege on Subject Matter and Identity Statements". *Analysis* 75 (4): 562–65. DOI: 10.1093/analys/anv073

Dickie, Imogen. 2008. "Informative Identities in the Begriffsschrift and 'On Sense and Reference'". *Canadian Journal of Philosophy* 38 (2): 269–88. https://doi.org/10.1353/cjp.0.0015

Frege, Gottlob. 1879. *Begriffsschrift, Eine der Aritmetischen Nachgebildete Formalsphrache des Reinen Denkens*. Halle: Nerbert

Frege, Gottlob. 1879a. *Conceptual Notation and Related Articles. Translated and edited by Terrel Ward Bynum*. Oxford: Clarendon Press

Frege, Gottlob. 1892. "Über Sinn und Bedeutung"/"Sense and Meaning". In Gottlob Frege. 1952. Translations from the Philosophical Writings of G. Frege. In Peter Geach & Max Black (eds.) Oxford: B. Blackwell: 56–78

Geach, Peter & Black, Max. 1952. *Translations from the Philosophical Writings of Gottlob Frege*. Oxford: Basil Blackwell. https://b-ok.cc/book/639351/b28968

Korta, Kepa. 2007. "Acerca del Monopropositionalismo Imperante en Semantica y Pragmatica". *Revista de Filosofía* 32 (2): 37–55. ISSN: 0034-8244

Korta, Kepa & Perry, John. 2011. *Critical Pragmatics*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511994869

May, Robert. 2001. "Frege on Identity Statements". In Carlo Cecchetto, Gennaro Chierchia, Maria-Teresa Guasti (eds.). *Semantic Interfaces: Reference,*

Anaphora, and Aspects. Palo Alto: CSLI Publications: 1–51. DOI: 10.1017/S0022226702221970

May, Robert. 2012. "What Frege's Theory of Identity is Not". *Thought* 1 (1): 41–8. DOI: 1002/tht.6

Mendelsohn, Richard. 1982. "Frege's Begriffsschrift Theory of Identity". *Journal of the History of Philosophy* 20 (3): 279–99. DOI: 10.1353/hph.1982.0029

Neale, Stephen. 1999. "Coloring and Composition". In Kumiko Murasugi & Robert Stainton (eds.). *Philosophy and Linguistics*. Boulder CO: Westview Press: 35–82. ISBN-10: 0813390850; ISBN-13: 978-0813390857

Perry, John. 1988. "Cognitive Significance and New Theories of Reference". *Nous* 22: 1–18. Reprinted in John Perry 2000. *The Problem of the Essential Indexical and Other Essays*. Palo Alto CA: CSLI Publications: 189–206. https://doi.org/10.2307/2215544

Perry, John. 2001/12. *Reference and Reflexivity* (2nd Edition). Palo Alto: CSLI Publications. ISBN 1-57586-433-9

Perry, John. 2020. *Frege's Detour*. Oxford: Oxford University Press. ISBN 978–0–19–881282–1

Simon, Peter 1995. "The Next Best Thing to Sense in Begriffsschrift". In John Biro & Petr Kotatko (eds.). *Frege: Sense and Reference One Hundred Years Later*. Dordrecht: Kluwer Academic Publishers: 129–40. ISBN: 978-94-011-0411-1

Thau, Michael & Caplan, Ben. 2001. "What's Puzzling Gottlob Frege?". *Canadian Journal of Philosophy* 31 (2): 159–200. https://doi.org/10.1080/00455091.2001.10717564

Organon F 29 (1) 2022: 47–77
https://doi.org/10.31577/orgf.2022.29103
ISSN 2585-7150 (online)
ISSN 1335-0668 (print)

RESEARCH ARTICLE

A Causal-Mentalist View of Propositions

Jeremiah Joven B. Joaquin* – James Franklin**

Received: 3 May 2020 / Revised: 20 September 2020 / Accepted: 24 January 2021

Abstract: In order to fulfil their essential roles as the bearers of truth and the relata of logical relations, propositions must be public and shareable. That requirement has favoured Platonist and other non-mental views of them, despite the well-known problems of Platonism in general. Views that propositions are mental entities have correspondingly fallen out of favour, as they have difficulty in explaining how propositions could have shareable, objective properties. We revive a mentalist view of propositions, inspired by Artificial Intelligence work on perceptual algorithms, which shows how perception causes persistent mental entities with shareable properties that allow them to fulfil the traditional roles of (one core kind of) propositions. The clustering algorithms implemented in perception produce outputs which are (implicit) atomic propositions in different minds. Coordination of them across minds proceeds by game-theoretic processes of communication. The account does not rely on any unexplained notions such as mental

* De La Salle University
http://orcid.org/0000-0002-8621-6413
Department of Philosophy, De La Salle University, 2401 Taft Avenue, Malate, Manila, Philippines 0922
jeremiah.joaquin@dlsu.edu.ph

** University of New South Wales
https://orcid.org/0000-0002-4603-1406
School of Mathematics and Statistics, University of New South Wales, Sydney, Australia 2052
j.franklin@unsw.edu.au

content, representation, or correspondence (although those notions are applicable in philosophical analysis of the result).

Keywords: Propositions; causal-mentalist view; cluster analysis; game theory; perception algorithms; Platonism; symbol grounding.

1. Introduction

Propositions play several theoretical roles. They are thought of, for example, as the primary bearers of truth and falsity, the *relata* of logical relations such as entailment, the objects of certain intentional states such as belief, and contents of linguistic acts such as assertion (McGrath and Frank 2018; Briggs and Jago 2012). There is disagreement, however, as to the nature of the posited entity that performs these roles. Platonist, possible-worlds, deflationary, and naturalist accounts compete.

We argue for one form of naturalist theory of propositions. In our view, propositions are non-abstract, structured entities, and we identify them with certain types of persistent mental entities created and coordinated by cognitive algorithms common to different minds. Work in Artificial Intelligence has discovered the kind of algorithms needed to create such mental entities, while the causal story of how they are created and work in game theory has shown how such entities can be coordinated across different minds.

We take no particular position on the nature of the mental, such as whether it is reducible to the physical. The mental is simply identified as whatever in humans plays the usual roles of intentionality, content, thought, perception and direction of action.

2. Directions for a theory of propositions

In this section we list several specifications which our theory of propositions will attempt to meet. Some are uncontroversial and some less so. We give brief reasons in their favour but cannot consider all the arguments of those who have taken different paths. Our aim is to motivate a causal-mentalist approach and to situate it in the spectrum of theories of propositions.

- **A theory of propositions will offer a clear answer as to the metaphyical nature of whatever is identified as propositions**

A theory should be clear as to whether it takes propositions to be Platonist abstract entities and if so which ones, or types of sentences and if so how the tokens are united into types (since a type is only identified by the properties that unite it), or mental entities and if so which ones exactly, or is a deflationary theory and if so what is left after deflation, and so on.

To defend a non-deflationary, or realist, theory of propositions is not necessarily to posit entities which are primary truth-bearers and to which thought or predication has some (external) relation. That way of speaking could concede too much to the Fregean Platonist view of propositions that once dominated the field. What the alternative might be, however, needs to become clear in the outcome. It will need to be explained how the entities to be posited accomplish the roles which propositions traditionally play.

- **It will explain both the public, objective aspect and the mental aspect of propositions**

Propositions are shareable and public in that two or more people can accept, believe, assert or communicate the same proposition. That is the point of them. Any account of the nature of propositions should attempt to explain those facts. *Prima facie*, a mentalist account of propositions could find this difficult, as mental entities are not public, nor do they appear to be shareable.

On the other hand, 'any [view] according to which propositions represent things as being a certain way and so have truth conditions in virtue of their very natures and independently of minds and languages is in the end completely mysterious and so unacceptable' (King 2009, 261; Pickel 2017). The problem of what propositions *are* does not arise from questions in physical or biological science. The roles played by propositions arise from humans attributing meaning to language, from language inducing thoughts, and from similar pieces of language reliably inducing similar thoughts and behaviours in different persons. Those *explananda* essentially involve the mental, so a theory of propositions must explain the relations of propositions to the mental. That is a *prima facie* difficulty for theories that propositions are to be identified with non-mental entities such as Platonist *abstracta*, sets of possible worlds, states of affairs or sentences. Our approach is thus

in accordance with the naturalist trends in cognitive science that have tended to supplant Fregean anti-psychologism in the last hundred years. We comment on Frege below.

- **It will explain compositionality: how the nature and roles of a proposition are a function of (or at least relate to) the nature and roles of its terms**

A well-formed sentence, like 'John is tall', surely expresses some thought. An ill-formed string of symbols, like '!@#$%', does not. Something must account for this difference. Likewise, something must account for the fact that 'John is tall' resembles 'John is short' in one way and 'Sue is tall' in a different way. The structural features of propositions account for the difference (Duncan 2018).

To think of propositions as structured entities is to think of them as complex entities having an ordered relation among their parts or constituents. The ordered relation of the proposition's constituents greatly matters not only in distinguishing between meaningful strings of symbols and nonsensical strings, but also in identifying whether two propositions with the same constituents are the same or different propositions. For example, the English sentence 'John loves Mary' expresses the same proposition as the Tagalog 'Mahal ni John si Mary', since structurally both have the same ordered set <loving, John, Mary>. But the sentence 'Mary loves John' expresses a different proposition from these two, since the proposition it expresses has a different structural ordering. Various accounts of the ordering relation are found in Schiffer (2012); King (2019); McGrath and Frank (2018); Hanks (2009).

While it is the majority position that propositions are structured, the great diversity in accounts of the nature of their parts opens a window onto what an account of propositions must explain (Briggs and Jago 2012; Chalmers 2012). Are propositions made up of symbols? Names? *Abstracta*? *Possibilia*? Intensions or meanings (King 2019)? Mental entities of some kind? Whichever is correct, surely any theory of the nature of propositions must include a theory of the nature of the parts and of the relation of parts to the whole.

- **It will explain the apparently close relation between propositions and states of affairs**

Propositions cannot *be* states of affairs such as 'Snow's being white' because there are no false states of affairs**.** It certainly seems that true and false propositions do not differ in their nature as propositions but only in their relation to how the world is. States of affairs are ways the world actually is and lack the mental or interpretive aspect of propositions. Nevertheless, it is surely not a coincidence that true propositions like 'Snow is white' have a structure that mirrors that of the corresponding state of affairs (and the structure of a false proposition can mirror that of a possible but non-actual state of affairs). It is desirable to have some explanation of that.

- **It will accomplish those tasks without reliance on philosophical notions which threaten to be as obscure than the *explanandum***

We aim for something more ambitious than accounts of propositions that take for granted concepts like 'representation', 'content of concepts', 'reference', 'intentionality', 'correspondence', 'information', 'third realm' or 'object of'. The problem is that those notions span the mental and the extramental in ways as mysterious as propositions themselves.

Leading contemporary mentalist theories of propositions have not attempted to do that. Soames's theory of propositions as 'cognitive event types' (Soames 2010) relies on a primitive notion of representation (Caplan 2016). Hanks's theory grounded on primitive 'acts of predication' (Hanks 2015) and Davis's theory grounded on 'declarative thought-types' (Davis 2005, 20–23) suffer this 'ungroundedness' problem.

We do not claim there is anything wrong with those notions, as philosophical interpretations of a causal story. Indeed, we will argue that propositions as we understand them do represent and their terms do refer. The objection is to taking those notions as primitive. We aim to give a free-standing causal account of the mental entities that play the role of propositions, with those entities arising in a way that does not rely on those notions. They can then be subject to an external philosophical analysis using notions such as representation and content.

- **Platonism will be a philosophy of last resort**

For Fregeans, the proposition expressed by a sentence is a structured relation of *senses*, where these senses correspond to the constituents of the

sentence expressing them. 'Senses' are not mental entities (a thesis dubbed 'psychologism' by Frege) but *abstracta*—the 'realm-mates of Platonic properties and relations' (Jubien 2001, 47; a clear account of the Fregean picture is found in Hanks 2015, 12–20). 'Abstract' here means 'abstract object' in the sense of full-blooded Platonism: non-spatial and causally inefficacious, as are numbers on a Platonist view (Rosen 2020). But Platonism in general labours under the weight of over two thousand years' worth of substantial objections such as its incompatibility with naturalism and its difficulties with epistemic access. In the context of propositions in particular, Frege's notion of their 'graspability' by the mind was always recognised as difficult and was a prime focus of attack by leaders of the (naturalist) 'cognitive turn' in analytic philosophy. (Sacchi 2006)

Platonism is a metaphysics of last resort—understandable if still problematic with apparently eternal objects like numbers, but surely even less attractive concerning entities like meanings that are so closely related to human utterances and intentions.

- **A naturalist, anti-Platonist theory of propositions will make essential reference to the causal origins of propositions**

Just as biology is incomprehensible without evolution, and a naturalist philosophy of mathematics would require explanation of how mathematical knowledge can arise in minds, so a non-Platonist theory of propositions requires an account of how they come to be. If propositions are not *abstracta* or other entities 'out there' and they play some role in cognition, they must fit into a causal, scientific story. Thus, we regard purely philosophical mentalist accounts of propositions such as Davis (2005), Soames (2010) and Hanks (2015) as incomplete though valuable. Instead, we will connect our theory with causal theories of reference (Devitt and Sterelny 1999), symbol grounding in Artificial Intelligence (Harnad 1990) and perceptual symbols in cognitive science (Barsalou 1999). We will argue however that by keeping to the correct level of analysis we do not need to delve into details of neuroscience.

- **A naturalist, anti-Platonist theory of propositions will maintain continuity between human and higher animal belief**

Although we cannot defend it at length here, our approach accepts that (some) propositional attitudes are in the first instance pre-linguistic, and so exist in higher animals and human neonates. As Fodor writes, 'You can, surely, believe that it's raining even if you don't speak any language at all. To say this is to say that at least some human cognitive psychology generalizes to infra-human organisms; if it didn't, we would find the behavior of animals utterly bewildering, which we don't.' (Fodor 1978, 512; at length in Nelson 1983) The basic causal mechanisms we discuss will be shared with higher animals.

- **Finally, the theory will then show how the entities identified as propositions fulfil the traditional roles of propositions as truth-bearers, objects of belief and other propositional attitudes, and *relata* of logical relations.**

This is the main task of any theory that purports to account for the nature of propositions.

To summarise the discussion so far and prepare for the next section, we present a classification diagram of theories of propositions. (Fig 1) Our theory belongs in the shaded box at bottom right

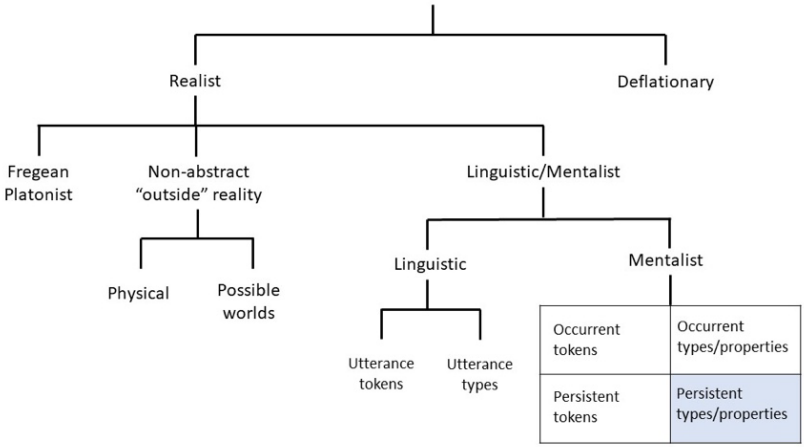

Figure 1: Classification of theories of propositions

The diagram will be referred to later in explaining the categories of the mental entities being discussed.

3. Propositions as mental entities?

So, why believe that propositions are mental entities? Soames notes that there is something unsatisfactory about the project of trying to place propositions 'out there' away from minds. He writes,

> The key to solving [the problem] is to recognize the obvious fact that predication is something that agents do… Instead of explaining the intentionality of the cognitive activity of agents in terms of an imagined conceptually prior intentionality of the propositions they entertain, we must explain the intentionality of propositions in terms of the genuine conceptually prior intentionality of the cognitive activity of agents who entertain them (King, Soames, and Speaks 2014, 33).

This agent-, mind-dependent nature of propositions is something that Platonist and other theories such as physicalist and possible-worlds theories have not accounted for satisfactorily. Propositions are bearers of meanings, and there is no getting past the fact that meaning is something intentional; something with a social aspect, certainly, but mediated through individual minds.

The mentalist option should therefore be revisited, despite the *prima facie* problems it faces. The dialectic for our view is simple. Since propositions are not 'out there', then they must be at least partly 'in here', or mental. Furthermore, since propositional attitudes like belief are mental, it would simplify matters if propositions were as well—the Platonist problems with relations between minds and *abstracta* would be avoided if propositions were 'inside' the mind.

But there is an obvious problem. Can we explain how two people can believe the same proposition? If a proposition is just in one individual's mind, as a token mental entity, then the proposition is not shareable. But the point was that propositions are shareable. Two people can both believe that Great Britain is a monarchy. Mental entities are private; propositions are, if not exactly on public display, at least interpersonally available.

On the other hand, while mental *tokens* are private, mental *types* can be shared. As such, these mental types can thus be treated as scientifically acceptable entities. To be more precise, the properties defining the type are shareable, while the tokens of the type are not. That possibility opens an opportunity for progress.

In order to maintain clarity about the nature of the mental entities to be discussed, we begin with several necessary distinctions, as laid out in the bottom right of Fig 1. A mental entity, such as a belief, may be either occurrent (a 'datable mental act', Textor 2017) or persistent (an underlying accessible mental structure that is activated as required, as when one is said to believe Pythagoras's Theorem even when one is not thinking about it.) The nature of mental life is for occurrent mental acts to be caused by some process such as perception, then persistent traces to be laid down in a process of learning.

A mental entity may also be token or type: my occurrent enjoyment of warmth is of a type with yours, while my persistent love of summer is also of a type with yours. It is important that the occurrent/persistent distinction is not the same as the token/type distinction. (It is also possible to distinguish between type and property—a species of birds is a type defined by possessing the properties characteristic of that species (Wetzel 2018, sec. 3)—but that distinction does not play an important role here.)

Should a mentalist theory aim to analyse persistent or occurrent propositions? Existing mentalist theories have preferred occurrent propositions. Soames (2010) uses mental *event* types, Hanks (2015, 6) similarly takes propositions to be types of 'mental and spoken actions', Davis's (2005, 20–23) 'declarative thought-types' are occurrent ones. We think, in accordance with the causal story to be told below, that long-term, persistent or dispositional, propositions (bottom right of Fig 1, shaded) are primary and occurrent thoughts or expressions of them are secondary and transitory. Propositions, like the terms such as names that are their constituents, should be persistent entities in the first instance which account for the unity of the event types that express them; so, their expressions in mental *events* such as predications are secondary.

However, we do not deny that the entities in the three other quadrants exist, nor that they can be rightly said to be true, or believed; one can

occurrently believe or speak truly. We will indeed defend a mentalist analogue of Hanks' claim that 'linguistic types inherit their semantic properties from their tokens.' (Hanks 2011, 41) That is natural since we begin with a causal story, and causes produce tokens—but ones with reliable properties since by and large like causes produce like effects, and it is those properties that account for their fulfilling the role of propositions.

4. Propositions as mental entities coordinated by cognitive algorithms

Propositions should be a kind of thing satisfying barely compatible requirements. They should be in some sense mentally dependent, but in order to be shareable they should not be individual (token) thoughts. They should be objective, yet neither denizens of the world of physical states of affairs nor of a world of *abstracta*.

4.1 Propositions as types of persistent mental entities

One possibility for the kind of thing that could in principle satisfy those requirements is a *property* of mental entities (and hence the type defined by the property). Properties, although sometimes called 'abstract', are not abstract objects in a Platonic realm, lacking causal power. On the contrary, things act in virtue of the properties they have, as when we see something as yellow because yellow things affect us in a certain way. One might well take an Aristotelian (anti-Platonist) realist view of properties (Armstrong 1978), but here we do not commit ourselves to any particular metaphysics of properties. We need only the scientific acceptability of causal properties and their powers, as when we observe that Newton's law of gravitation relates (the properties) gravitation, mass and distance, or Weber's law relates stimulus (in general) to perception (in general). Properties as we discuss them will be understood in that minimal scientific and naturalist sense, according to which particulars have causal powers in virtue of the properties they have.

Strangely, Frege's own arguments for his Platonist position do not rule out this possibility. He argues concerning the 'thought' or what we would call the content of a proposition:

> Is a thought an idea? If the thought I express in the Pythagorean theorem can be recognized by others just as much as by me then it does not belong to the content of my consciousness, I am not its bearer … If every thought requires a bearer, to the contents of whose consciousness it belongs, then it would be a thought of this bearer only and there would be no science common to many … So the result seems to be: thoughts are neither things of the outer world nor ideas. A third realm must be recognized (Frege 1956, 301–2).

That is correct, but to conclude that if something is neither a physical object nor an occurrent idea, then it must be a Platonic abstract object neglects certain other possibilities. Properties of mental entities are another, and, as we have explained, they are not *abstracta* (as they have causal powers, like the properties of physical objects). Properties of mental entities satisfy Frege's desiderata for 'thoughts', that they be apprehensible by many minds because they belong 'neither to my inner world as an [occurrent, particular] idea nor yet to the outer world of material, perceptible things' (Hanks 2015, 3–4).

My (occurrent) thought of cats and your thought of cats are numerically different mental entities, one in my mind and one in yours. But they can have properties in common. For example, they can occur at the same time, and they can both be partially caused by perceptual experience of cats. So, in principle, my thought and your thought could have a property (or stand in a relation) that makes them (tokens of) the same proposition. But what property, and how is that property acquired? It needs to be a property anchored in the 'out there' (like the causation of perception by real cats), that gives the two thoughts, yours and mine, in some way a sufficiently similar relation to external states of affairs. 'Out there' should mean, in naturalist fashion, out there in the real physical and social world, not out there in a Fregean Platonic 'third realm' causally detached from the real physical and social world.

A beginning on explaining how this is possible is made by a propositional theory of perception, such as that of Armstrong's *Perception and the Physical World.* Armstrong writes:

> Physical objects or happenings stimulate our sense organs. As a causal result of this we acquire immediate knowledge of their existence and their properties … This knowledge is not necessarily verbalized knowledge, but it is always knowledge which it is logically possible to put verbally. It is propositional in form... The acquiring of immediate knowledge in this way is *perception* (Armstrong 1961, 191).

That is of no immediate help in analysing what propositions are: since it analyses perception in terms of propositions, it leaves propositions themselves unanalysed. But it does suggest that, if some independent account can be given of how perception creates mental entities, the entities so created can fulfil the role of propositions (at least, those propositions stating claims about perceived reality). In particular, since everyone's perceptual apparatus is set up by biological causes to be similar, there will normally result a high degree of commonality between the results of my perception and yours (of the same thing). If the results of perception are propositions, then it will have been explained how the mental entities common to our believing a proposition do have properties in common: they are caused by the same perceptual object affecting the mind through a similar causal process.

4.2 Cluster analysis to generate atomic perceptual propositions

We begin by separating the case of propositions about immediate perceptual reality from others, such as inferred propositions and ones believed on the basis of testimony. That is in accordance with the long philosophical tradition from the Aristotelian 'nothing in the intellect that was not first in the senses' through Locke to Carnap's *Aufbau*. It is also in accordance with much recent philosophy of language. As Devitt and Sterelny put it,

> A causal theory of natural kind terms, like one for names, divides in two. First, there must be a theory of reference fixing, which explains how a term is linked to a referent in the first place … Second, there must be a theory of reference borrowing, which explains the social transmission of a term to those having no contact with its referent (Devitt and Sterelny 1999, 88).

That applies to propositions as much as to the terms in them. There are good reasons for that separation into two levels. The understandability of 'It's sunny in California' is parasitic on the understandability of 'It's sunny here'. If we can grasp the nature of immediate perceptual propositions, we can hope to move on to others which stand in a logically dependent relation to them (such as logical combinations of immediate perceptual propositions, or merely possible ones, or reported ones).

The two-level approach is also suggested by the commonality of perceptual recognition to us and nonhuman primates. The ability to form presymbols out of perceptual input is found in the higher primates, but apparently not the compositionality and inference of human language (Zuberbühler 2018), nor reference to more distant or abstract entities. As Wittgenstein neatly puts the two stages, 'A dog believes his master is at the door. But can he also believe his master will come the day after tomorrow?' (Wittgenstein 1953, 174; discussion in DeGrazia 1994) Therefore, it would be desirable if the most basic account of what a proposition-like entity is could rely only on what is possible for the primate mind, leaving more complex operations peculiar to humans for a later stage.

Certainly, achieving an independent account of perceptual grounding has proved very difficult, both at the in-principle level desired by philosophy and at the more detailed level required by cognitive science (and even more so, at the implementable level sought by Artificial Intelligence). The extensive work in those areas has however produced some models that are useful in demonstrating how a mental entity such as one whose properties play the role of a term or proposition could be at once internal to the mind, caused by an appropriate part of physical reality, and endowed with properties shareable by other minds. We will use the simple example of cluster analysis as applied to the symbol grounding problem (Franklin 1996). This is not simply an example or metaphor, but an exercise in abstract task analysis to identify what must be done, minimally, to create an atomic perceptual proposition.

Corresponding to the old philosophical problem of how words get their meaning, cognitive scientists address the 'symbol grounding problem' (Harnad 1990). In an Artificial Intelligence system intended to perform tasks like computer vision, how can the internal symbols which are to represent

pieces of external reality actually be attached to and caused by the perceptual input from that reality? In the vast flux of changing pixel values input to a vision system, how is the relevant part to be identified and labelled so that the system locates and tracks an object? The problem is not simply a technical one: here the main interest is in the task specification and in understanding the kind of algorithm that could possibly solve it, along with the nature of the result. (The distinction between task specification and algorithm is that a task specification states what output is to be created from what input—'make a chocolate cake from these ingredients'; 'sort a list of words into alphabetical order'—while a recipe or algorithm lists the steps to accomplish the task—a definite cake recipe for the cake; for sorting, an algorithm such as bubble sort which repeatedly goes through the list and swaps adjacent items if they are in the wrong order.)

Cluster analysis works like this. Its purpose is to take a heap of points as in Fig. 2 (that is, it is just given unlabelled points with their positions), and to conclude 'There are two clusters, and these points are in cluster A and those are in cluster B.' To the eye, it is easy, but finding and programming an algorithm that performs the task is non-trivial.

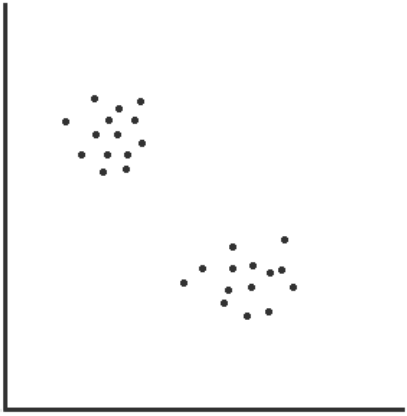

Figure 2: Points in two clusters[1]

[1] Figure adapted from: http://www.philender.com/courses/multivariate/notes2/cluster0.html.

The main application of cluster analysis is to 'points' that are not in geometrical space but in 'feature space' (as in Fig. 3, but typically the space has many dimensions). The axes represent features of objects and the degree to which objects possess them. So, a dot placed in the space represents an object, with its position in the space representing the degree to which it has each feature.[2]

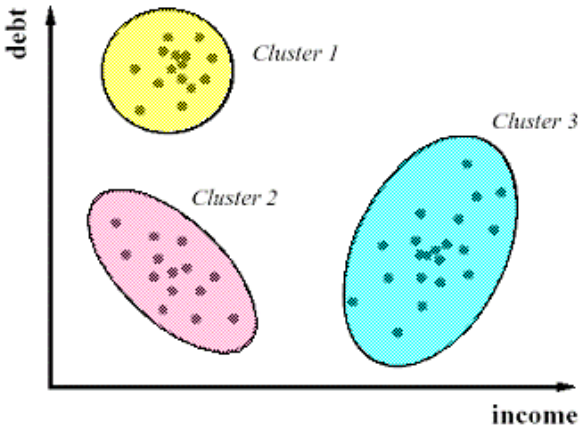

Figure 3: Clusters of points in feature space[3]

For a vision system (artificial, animal or human) to recognise an object against background requires a form of cluster analysis: the pixels in the object, which are similar to one another in both position and colour, need to be 'stuck together' by the system's software to create the 'object' cluster, while being separated from the 'background' cluster of pixels.

Then once one has an individual cat (say) cut out of the background and identified as a single object, one must perform cluster analysis again (in feature space) to recognise that cats are similar to one another and dogs are similar to one another (across a range of features) and there is not much in between. Hence, there are recognisable natural kinds: cat and dog.

[2] For a technical survey, see (Jain, Murty, and Flynn 1999).

[3] Figure adapted from: http://2.bp.blogspot.com/_CWYkOgzhyq0/TAJ4oFopasI/AAAAAAAAADo/NRrS4E3R1cs/s400/cluster_analysis_income_debt.gif.

How similar inputs have to be (and similar in what ways) to count as in the same cluster is something that the clustering algorithm has to work out for itself. Cluster analysis is an exercise in discretisation—different points in a continuous input space end up classified into a single discrete cluster. Whether different perceivers can recognise sufficiently similar clusters given their somewhat different perceptual inputs is a matter for scientific investigation. The fact that communication using perceptually-generated concepts often succeeds suggests that they do. It is easy to think of the difficulties of classifying handwritten postcode digits, but software exists which does almost perfectly classify them into ten classes.

It is not merely that cluster analysis might be helpful for the problems of early perceptual grouping and of symbol grounding, but that the nature of the problems means that *any* solution to them must be some form of cluster analysis. Such problems all involve forming a discrete object out of a cluster, that is, a mass of neighbouring data points that are all close among themselves, but are all reasonably well separated from other data points.

A clustering algorithm is a perfectly comprehensible series of steps, a recipe implementable in software—whether software written by humans in an AI system or the mental implementation of such an algorithm in the human mind, however that is accomplished (and because we are working at the level of task specifications and algorithms, the theory can remain neutral on the mechanisms by which that is in fact accomplished). An implemented algorithm is a naturalistically acceptable entity, not an inhabitant of a Platonic world: it is just a series of regular scientifically-discoverable steps with inputs and outputs, like the process of photosynthesis. The algorithm does not contain any philosophical overhead such as 'representation' or 'content'.

Nevertheless, its output consists of discrete items that it would be natural for the outside observer to recognise as uncannily resembling terms and propositions. The clusters identified by the algorithm in the flow of experience are labels of naturally grouped persistent objects. While the algorithm requires some structured input—the inflow of perceptual input must have some inhomogeneities actually present, and the feature dimensions are also given—the objects have been discovered in the data by the algorithm, not

presented to it beforehand. Nevertheless, we can find in its output items bearing a resemblance to all of the constituents of an atomic proposition *Fa*. The 'subject' *a* is a cluster found in the points (the pixel positions and values, for example, clumped by the cluster algorithm to form natural discrete objects such as spot against background). The predicate *F* is a region of feature space; for example, the bottom left cluster in Fig. 3 lies in the low-income region of space. (That does require that the system has the capability to represent explicitly dimensions of feature space and their parts; we do not attempt to solve that problem here.) The outside observer is welcome to use the language of representation, so as to agree with Fodor's 'there are internal representations and propositional attitudes are relations that we bear to them.' (Fodor 1978, 519) But representation is an interpretation of what has arisen from the algorithm, not something assumed or input at the start.

The proposition is the (persistent type) association between cluster and region, that is, the output of the clustering algorithm identifying the cluster and where it lies in feature space. One must distinguish between the points in the cluster actually being in that region (a state of affairs in the real world), and the clustering software's explicit identification of the cluster as being in that region (an output of the software). It is the latter that is a proposition. As it is a type, produced similarly in different correctly-functioning minds, it can be common to different minds. In a pre-linguistic animal, there is no more to belief in a perceptual proposition than the algorithmic output (which should be sufficient for action on the basis of that belief). Humans may, as is their wont, add some conscious thoughts and linguistic expressions, but those are not essential to the belief in the proposition.

As with any implemented software, a system performing clustering can be analysed at three levels of abstraction. At the lowest level is the working implementation (code running on a machine, in the computer case; working neurobiology, in the human case). At the next level is the algorithm: the recipe or sequence of steps that is implemented in the code or neural activity. At the highest level is the task analysis or program specification, which describes what the algorithm is to do (in terms of transforming inputs to outputs) without laying down the steps it is to perform to do it. Different algorithms (such as different clustering algorithms) may perform the same

task (such as identifying natural clusters in data). For the purpose of identifying propositions, the important level of analysis is the highest, that of specifying the clusters (of objects and kinds) formed by the perceptual clustering. Thus, if Martians have a different algorithm for perception, but identify the same clusters in data, such as cats and mats, their proposition 'the cat sat on the mat' is the same proposition as ours, in virtue of their algorithm satisfying the same program specification as ours.[4]

The same—task analysis—level is appropriate for deciding whether two people believe the same proposition, such as 'the cat sat on the mat,' on the basis of their slightly differing perceptual experiences. The aim of perceptual algorithms is exactly to create persistent discrete unities out of the continuous flux of perceptual experience. It may succeed exactly, approximately, or not at all, and whether it does so the same way in two perceivers is a scientific question to be examined in the usual ways, such as by checking how similar are their inferences, answers and behaviour on the basis of the proposition believed.

Although task analysis or program specification is an obvious sense 'more abstract' than lower levels, it does not follow that it involves any naturalistically unacceptable *abstracta*. Just as photosynthesis has a more efficient implementation pathway in tropical plants than in temperate-zone ones but still effects the same biochemical transformation (without calling on any non-naturalist entities), so different cognitive algorithms may perform the same cognitive task, such as creating atomic perceptual propositions.

4.3 Properties of perceptual propositions generated by cluster analysis

It is not difficult to read off answers, in that model, to some of the troubling questions about propositions described earlier—at least for atomic propositions that summarise perceptual input.

As to the metaphysical nature of a proposition, it is an explicit internal (mental) persistent discrete entity, generated by internal algorithms acting

[4] The separation of levels of analysis is less clear in the 'perceptual symbol systems' of Barsalou (1999), which otherwise attempts a similar causal analysis to the present one.

(in this case) on perceptual input. 'Explicit' does not mean 'conscious'—the mind's insight into its own workings is notoriously weak—but instead means a discrete output of mental processing, capable of entering into further processing, and in principle capable of being identified by the sciences of psychology and neuroscience.

The way in which the mental (persistent) tokens of propositions in different minds are coordinated and share properties follows from their being the outputs of identical (or very similar) algorithms run on similar data. Just as your cake and mine resemble each other (have similar persistent properties) if we use the same recipe on similar ingredients, so your token internal proposition that there is a cat before us resembles mine in virtue of our having run similar perception-processing clustering algorithms on similar visual input. It is true that if your experience of cats is only of white ones and mine only of black ones, there may be some mismatch between our concepts and thus potential for miscommunication. That is inevitable and algorithms can only do so much with the data they are given.

Compositionality has been explained above: the terms *a* and *F* in the proposition *Fa* are themselves explicit as outputs of the clustering algorithm—they are exactly what the algorithm is designed to output. So is their relationship in the proposition. The fact that terms and propositions are discrete outputs is what allows them to become inputs in the higher levels of discrete processing: for example, to enter into explicit conjunctions and other logical composites and to be expressed in discretely structured natural languages. At least, it allows that in principle. There are some difficult questions as to how human cognition, but apparently animal cognition only minimally, achieves some explicit knowledge of the representational nature of its internal symbols. Those questions are about cognition and not directly about the nature of propositions and cannot be addressed here.

The relation between proposition and state of affairs is, in this simplest case though not necessarily elsewhere, the causal one between data input and algorithmic output. The cat's being on the mat is a state of affairs.[5] The perceptual clustering algorithm identifies discrete objects *cat* and *mat*

[5] We take a realist approach to states of affairs as defended in (Armstrong 1997) and regard them as unproblematic entities in the present context, and as composed of particulars and universals as they seem to be.

in the visual experiential flux, tagged with their positions in space, computed from the visual directions of their parts. Some further processing, but of the same nature, is required to make explicit the spatial relation of cat and mat, inherited from the spatial relations of the perceived pixels belonging to each.

There is more than one way in which that simple causal relation of proposition and state of affairs can become more complicated. Firstly, there are many possible ways to go wrong between input and output; any algorithm implemented in real software can malfunction and, for example, output cluster labels when there are really no clusters in the data. In that case, there is no state of affairs corresponding to the output. Correspondence is thus a relation that may or may not hold between states of affairs and the output of the clustering software: there are certain clusters really in the data, and the algorithm does or does not successfully find them. No external notion of truth has been invoked—correspondence is defined in terms of the properties of the data and the software's performance on the data. Again, the contrast between animal and human cognition is useful for understanding what has been claimed. An animal's internal proposition can rightly be said to represent external reality, but that is a comment made by an outside human philosophical observer and is unknown to the animal. Humans have more insight into their cognition (we do not claim to have explained how) and their awareness of the representational nature of their symbols is useful not only philosophically but for such purposes as logical inference.

Secondly, there are other possibilities for lack of correspondence—mismatches between reality and the outputs of the software — when the discrete outputs of clustering algorithms are used as the inputs of recombinations. Human mental capacities, though apparently not cockroach mental capacities, include explicit recombination of the discrete chunks. '*a* is red' and '*b* is blue' permit the recombination '*a* is blue'[6]; 'X loves Y' permits the recombination 'Y loves X'. This 'systematicity' of thought was adduced in Fodor and Pylyshyn's celebrated (1988) paper as being incompatible with a connectionist architecture for the mind; certainly, it points to the requirement that discrete recombinable entities should be found at a basic level in

[6] Provided that 'red' and 'blue' are recognised as in the same category, something which itself needs to be a capacity of the software.

thought. The ability to 'mix and match' items is essential in exploring 'what-if' scenarios—thought experiments where humans imagine what would happen in situations that have not occurred yet, or 'mental time travel' (Suddendorf and Corballis 2007). It is the foundation of the human ability to plan. Recombination of (categorically compatible) terms yields something that might be the output of the clustering algorithm applied to some possible state of affairs. Or again, it might not, if the recombination yields an impossible state of affairs: there is no guarantee that recombination of mental items tracks what is really compossible in the outside world. A third way in which the relation of proposition and state of affairs may become more complicated arises when the proposition is more logically complex than an atomic *Fa* or *Gxy*. Those ways cannot be described in detail here. An example is that the relation of the truth-functionally complex proposition '*Fa* and *Gy*' to states of affairs can be explained as inherited from the relations of *Fa* and of *Gy* to states of affairs; that is the point of the connective 'and'. Another story can be told about propositions with complexity like 'X believes that *p*', but clarification can be left until an account is given of how propositions fulfil roles such as objects of belief.

Not mentioned so far is any conscious mental content. It is not denied that there may be such qualia as feelings of *aboutness* or cognitive satisfaction when one entertains propositions. However, just as qualia of blue and red are not an essential part of the story of human perceptual discrimination between blue and red—which may take place without or before any such conscious sensation, even if such qualia exist—so the essential mental entity, the proposition, can perform its cognitive role whether or not accompanied by any qualia or awareness.

4.4 Propositions other than perceptual

It is of course not claimed that clustering algorithms applied to perception can perform an indefinitely large range of knowledge-generation tasks. Early perceptual grouping and object identification form a restricted range of knowledge, albeit a crucial and foundational one. Clustering is just a first and easily understood example, where the relation of input to output can be studied free of complications.

However, a good deal is known at least in principle about the generation of non-perceptual propositions from perceptual ones. In developmental psychology, for example, evidence on the production of more inferential knowledge, such as research on 'Bayesian babies', suggests that humans share inbuilt algorithms for inferential knowledge too (Denison, Reed, and Xu 2013). Something was said above about the human ability to speak of merely possible states of affairs through recombination. In the philosophy of language, much has been said about how the social nature of language (once the developing infant has inferred that a world of other minds exists out there) allows for one person to 'catch' propositions from another (Devitt and Sterelny 1999, 96–101). These abilities—apparently very minimal in animals—create a great variety of propositions. They require complex stories as to how they refer to reality (or fail to), but there is no reason to think they require a different account of the nature of propositions. If we understand the nature of perceptual propositions, recombinations of them and their parts are not particularly mysterious in principle.

Propositions involving reference to fictional and abstract entities also need their own story. Again, animals and probably human neonates seem unable to entertain such propositions, confirming the desirability of a two-stage approach that starts with perceptual propositions, and suggesting that such questions do not concern the nature of propositions but the nature of those entities. We agree with the argument of Moltmann (2013) that reference to abstract entities is both special and rare. No doubt the ability to refer to fictions and *abstracta* is a development of the ability to recombine. For example, the recognition that 'a is red' and 'b is blue' permit the recombination 'a is blue' can suggest detaching 'blue' as an entity in itself which can be the subject of propositions. If our theory satisfactorily covers the nature of propositions in animals and neonates, its extension to the range of sophisticated entities discussed by adult humans is a matter for special investigations on those topics.

As pointed out by McDaniel (2005), any realist theory of propositions also needs an account of propositions that have never been and will never be entertained by anyone. That is easy for Platonist theories—they all exist in the same way as entertained propositions—but harder for mentalist or linguistic theories. We argue that, as for the question of propositions about

abstract entities, it is a problem, but not one about the nature of propositions. It is a problem about the nature of uninstantiated properties. Just as the reality (or otherwise) of an uninstantiated shade of blue is a problem about the metaphysics of properties (Franklin 2015) rather than a problem in the science of colour, so unentertained propositions (on a mentalist view) is a problem about the uninstantiated in general rather than about propositions.

However, an analogy will explain why the problem of unentertained propositions should not be regarded as serious for a naturalist account. Let us consider the corpus of programs written in some computer language; for definiteness let us take an obsolete one such as FORTRAN V, in which a finite number of programs were written but which is no longer used. Suppose we are comparing two philosophies of the nature of computer programs. A Platonist one holds that computer programs (in all languages actual and possible) exist in a Platonic heaven, and an infinitesimal proportion of them are written down by some programmer. An alternative naturalist philosophy holds that computer programs are creations of programmers, even though it is an objective matter which of the symbol strings written down are well-formed FORTRAN V programs. A program is an actual string of symbols created by a person (or machine), and it is a well-formed FORTRAN V program if it follows certain rules. The Platonist will urge the "problem of unwritten FORTRAN V programs" as an objection to the naturalist theory. How serious is that objection?

It is not a serious objection. Many natural processes are generative of possibilities that are never realised. Darwinian evolution could generate many species other than those actually found in the history of life on earth, while organic chemistry could generate many compounds other than those it actually does. That is no reason to think that species or organic compounds are really Platonic entities. The Platonist does believe that any species or compound, actual or possible, reflects some Platonic archetype, but for the naturalist who rejects that on general metaphysical grounds, there is no further "problem of unrealised species/compounds/programs". Those are just latent (and predictable) possibilities of the natural generative process. It is the same with propositions: given a naturalist story of how entertained ones are formed, such as has been presented here, the process

has natural generative possibilities for further propositions that may not ever be entertained. But that is not a reason to reject a naturalist process.

The case of false propositions is considered below; false propositions, even perceptual ones, cannot be directly caused by perceivable reality.

5. Coordination of propositions between minds via game theory

Before turning to the question of how the mental entities that have been identified as propositions fulfil the roles initially laid down for propositions, such as being truth-bearers and objects of belief, one further issue needs to be addressed, namely the coordination of propositions between different believers (human believers, if not animal ones), and the public standing of propositions. Although it has been explained how (tokens of) propositions in different minds can share properties in virtue of being the output of similar algorithms on similar data, that still leaves propositions as (types of) private entities hidden in minds. How can they acquire sufficient public standing to allow communication? How can there be reliable coordination between propositions in different minds, so that standard communication cues like words and gestures reliably induce the same proposition (type) in different minds? That is a further task, which it seems that humans can do but animals cannot—cats can discretise perceptual input in a similar way to humans (that is a minimal requirement for recognising conspecifics and prey, which higher animals can do), but they do not appear to coordinate their internal propositions with other cats.

As a simple model for how that is possible, and possible for the sort of mental entities that have been identified as propositions by the theory being put forward here, consider the stability of strategies in game theory. In an Iterated Prisoner's Dilemma game, the tit-for-tat strategy is a stable equilibrium. The game has two players, with a simultaneous choice at each step between two moves, 'cooperate' and 'defect'. If both cooperate, the payoff is better for both than if both defect, but a temptation to defect results from the rule that if one player cooperates and the other defects, the defecting player is well rewarded and the cooperating player punished. The

tit-for-tat strategy (play the move that the other player played in the last round) permits cooperation to develop, to the benefit of both players, but avoids the perils of being played for a sucker time after time. The game is mentally dependent, in the sense that a round of play occurs as a result of the players' intentions to play it (and of each player's knowing that the other has such an intention). It is an objective, mind-independent fact, however, that tit-for-tat is the best way to play it—the way that leads reliably to the best outcome for each player in the medium to long term—so that rational players will tend towards that strategy. That strategy will then typically be observed to be implemented in the game (if the players are indeed rational): it will be observed that plays do agree with the move chosen by the other player in the previous round, and questioning may elicit agreement from the players that they are following that strategy. There is a mental type whose tokens in each player mean they are both playing tit-for-tat. Neither a pattern observed in the sequence of moves nor an intention to follow such a pattern is normally thought to require the existence of any Platonic entity such as an 'abstract strategy'. There are no entities metaphysically more mysterious than instantiated patterns and the intentions of individuals.

The analogy between game-theoretic strategies and the interpersonal coordination of propositions is closer than it looks. To speak minimally and somewhat naively: propositions can typically be expressed in language, and one main purpose of doing so is to communicate. Communication is, among other things, a game-theoretic exercise. When a speaker enunciates a sign with the intention that the hearer should read it as a sign and take it in a certain way, the hearer's move in the cooperative communication game involves his guessing the way in which the speaker intends the sign to be taken. It is an iterated game, so the history of cues that have worked is relevant to future moves. Thus, public language coordinates the induction of propositions across minds. (This is similar to the philosophy of communication of Lewis (1969), and Grice's (1989, chap. 2) 'cooperative' principle of conversation, needed to allow a hearer to infer what a speaker intends from the speaker's public utterance.) That does not imply, however, that the tokens of propositions themselves are public or somehow in the external world. Like the intention to play a round of a game, the token proposition

is a mental entity, but communication is possible because of the coordination of mental entities driven by the game of communication.

6. How propositions as algorithmically coordinated mental entities fulfil the traditional roles of propositions

It remains to be explained how this theory of propositions—mental entities sharing properties through their creation by common algorithms and coordinated by game-theoretic communication requirements—accounts for the roles of the proposition, as listed in the first section above.

Their role as *truth-bearers* follows from what has been said about correspondence. An implementation of an algorithm to find the real clusters in perceptual data can succeed or fail in doing so, whence the result can be called true or false. (Allowable) recombination of the discrete outputs of the algorithm—the terms of a potential proposition − creates an internal proposition that could be the output of the algorithm applied to some possible state of affairs; the recombination is true or false according as that state of affairs obtains or not.

Propositions as conceived here are the *objects of belief* because they are actually identified with (implicit, dispositional) beliefs. There is nothing to the implicit belief that p over and above its having been installed in the mind by the belief-creating algorithm; which in the simplest case is the clustering algorithm applied to perception. (Of course, inferred beliefs need a further account of the process creating them out of immediate beliefs, but that does not bear on the nature of propositions itself.) Explicit, occurrent beliefs, such as consciously entertained or linguistically expressed ones, are created by the mysterious process by which the human mind can reflect on and bring to consciousness some of its contents. How that happens is again not directly relevant to the nature of propositions.

The story about the origin of propositions does not involve language in any essential way, so it has been explained in principle how propositions can be the *meanings common to sentences* of different languages. The clustering algorithm is independent of language—indeed, it is a precondition of the learnability of language. For an infant to associate the sound ‘cat’ with

experiences of cats, it must have solved four perceptual discretatision or clustering problems first: (1) It has to cut individual cats out of the visual background, (2) it has to classify those objects as (potentially) a single kind, then (3) it has to segment the sound stream so that the syllable 'cat' is isolated, and finally, (4) it has to classify that occurrence of 'cat' with other occurrences of that syllable. Only after that can there be association between the syllable-type 'cat' and experience of the type cats. Thus, organising perceptual experience into discrete repeatable pieces has to take place before questions about the meaning of words can even arise. It is to be expected that linguistic expressions of simple terms and propositions would be intertranslatable between different languages, although there is scope for different languages to construct different complex concepts from combinations of the simples, resulting in some mismatch of meanings of statements in different languages. That is observed to be the case (Wierzbicka 1996).

It is harder to explain why propositions as conceived here, as coordinated mental entities, should be *the relata of logical relations*. It is not clear why mental entities, even when tied down and given objectivity by the causality of an algorithm or the coordination imposed by the cooperative game of communication, should be subject to the absolute, mind-independent necessity of relations such as logical consequence or logical contrariety. The clue lies in the connection between logical relations and truth. As an account of the truth of propositions has already been given, in terms of a certain kind of correspondence between propositions and states of affairs, logical relations should be explained in the same way. A particularly clear case is the close connection between the logical consequence relation between (certain) propositions and the inclusion relation between states of affairs. The state of affairs of this raven's being black is a *part* of the state of affairs of this raven and its neighbour both being black. It is thus impossible for the latter state of affairs to obtain without the former also obtaining. If now a perceptual clustering algorithm (correctly) applies to those states of affairs and extracts the propositions that this raven is black and that both ravens are black, respectively, then the relation of logical implication holds between the latter and the former. It 'mirrors', so to speak, the inclusion relation between the states of affairs, meaning that the inclusion relation between the states of affairs is the truthmaker of the fact that it is

impossible for the second proposition to be true without the first also being true. That is, the latter proposition implies the former.

While that is the start of the story, it does not cover all cases. It does not explain why the proposition that this raven is white is implied by the proposition that this raven and its neighbour are both white, because those propositions are false and so do not correspond to any states of affairs, hence there is no inclusion relation between states of affairs to act as the truthmaker of their logical relation. Nevertheless, the potential for those propositions to result from actual states of affairs determines their logical relations. This raven's being white and these two ravens' being white are possible states of affairs, the first of which would be a part of the second, if they obtained. Hence, it is impossible that the second should obtain without the first, whence the second proposition logically implies the first.

Naturally that will create a problem for propositions that purport to describe what are in fact impossible states of affairs. But again, it is not the business of the theory of propositions to explain the philosophical complexities of all the things that propositions could (or could not) be about.

7. Conclusion

There is every reason to agree with the tradition that there are such things as propositions, entities that fulfil roles such as objects of beliefs and relata of logical relations. But there is no reason why the requirement that propositions be objective and shareable should lead philosophy on a wild goose chase into a Platonic realm of *abstracta*. Objectivity is available to mental entities, via the cognitive algorithms and game-theoretic coordination strategies that we all share. Those algorithms induce in different minds (implicit) beliefs. The beliefs in different minds share properties that allow them to fill the roles of propositions.

Acknowledgments

A version of this paper was presented at the Propositions, Reference, and Meaning Workshop held in November 14-15, 2019 at the Institute of Philosophy, Slovak Academy of Sciences, Bratislava, Slovakia. We would like to thank the

organizers and participants of the workshop for their useful comments and suggestions.

References

Armstrong, David M. 1961. *Perception and the Physical World*. New York: Humanities Press.

Armstrong, David M. 1978. *Universals and Scientific Realism*. Cambridge: Cambridge University Press.

Armstrong, David M. 1997. *A World of States of Affairs*. Cambridge: Cambridge University Press.

Barsalou, Lawrence W. 1999. "Perceptual Symbol Systems." *Behavioral and Brain Sciences* 22 (4): 577–660. https://doi.org/10.1017/S0140525X99002149

Briggs, Rachael, and Mark Jago. 2012. "Propositions and Same-Saying: Introduction." *Synthese* 189 (1): 1–10. https://doi.org/10.1007/s11229-012-0091-1

Caplan, Ben. 2016. "Soames's New Conception of Propositions." *Philosophical Studies* 173 (9): 2533–49. https://doi.org/10.1007/s11098-016-0633-3

Chalmers, David. 2012. *Constructing the World*. Oxford: Oxford University Press.

Davis, Wayne A. 2005. *Nondescriptive Meaning and Reference: An Ideational Semantics*. Oxford University Press.

DeGrazia, David. 1994. "Wittgenstein and the Mental Life of Animals." *History of Philosophy Quarterly* 11 (1): 121–37.

Denison, Stephanie, Christie Reed, and Fei Xu. 2013. "The Emergence of Probabilistic Reasoning in Very Young Infants: Evidence from 4.5- and 6-Month-Olds." *Developmental Psychology* 49 (2): 243–49. https://doi.org/10.1037/a0028278

Devitt, Michael, and Kim Sterelny. 1999. *Language and Reality: An Introduction to the Philosophy of Language*. 2nd ed. Boston, Mass.: MIT Press.

Duncan, Matt. 2018. "Propositions Are Not Simple." *Philosophy and Phenomenological Research* 351–66. https://doi.org/10.1111/phpr.12362

Fodor, Jerry. 1978. "Propositional Attitudes." *Monist* 61: 501–23. https://doi.org/10.5840/monist197861444

Fodor, Jerry A., and Zenon W. Pylyshyn. 1988. "Connectionism and Cognitive Architecture: A Critical Analysis." *Cognition* 28 (1): 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Franklin, James. 1996. "How a Neural Net Grows Symbols." *Proceedings of the Seventh Australian Conference on Neural Networks*, Canberra, 91–6.

Franklin, James. 2015. "Uninstantiated Properties and Semi-Platonist Aristotelianism." *Review of Metaphysics* 69 (1): 25–45.

Frege, Gottlob. 1956. "The Thought: A Logical Inquiry." *Mind* 65 (259): 289–311.

Grice, H. P. 1989. *Studies in the Way of Words*. Princeton, NJ: Harvard University Press.

Hanks, Peter. 2009. "Recent Work on Propositions." *Philosophy Compass* 4 (3): 469–86.

Hanks, Peter. 2011. "Structured Propositions as Types." *Mind* 120 (477): 11–52.

Hanks, Peter. 2015. *Propositional Content*. Oxford: Oxford University Press.

Harnad, Stevan. 1990. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42 (1): 335–46. https://doi.org/10.1016/0167-2789(90)90087-6

Jain, A. K., M. N. Murty, and P. J. Flynn. 1999. "Data Clustering: A Review." *ACM Computing Surveys* 31 (3): 264–323. https://doi.org/10.1145/331499.331504

Jubien, Michael. 2001. "Propositions and the Objects of Thought." *Philosophical Studies* 104 (1): 47–62. https://doi.org/10.1023/A:1010361210072

King, Jeffrey C. 2009. "Questions of Unity." *Proceedings of the Aristotelian Society* 109 (1pt3): 257–77. https://doi.org/10.1111/j.1467-9264.2009.00267.x

King, Jeffrey C. 2019. "Structured Propositions." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/propositions-structured/

King, Jeffrey C., Scott Soames, and Jeff Speaks. 2014. *New Thinking About Propositions*. Oxford: Oxford University Press.

Lewis, David K. 1969. *Convention: A Philosophical Study*. Vol. 20. Oxford: Wiley-Blackwell.

McDaniel, Kristopher. 2005. Review of D.M. Armstrong, *Truth and Truthmakers*. *Notre Dame Philosophical Reviews*, https://ndpr.nd.edu/news/truth-and-truthmakers/

McGrath, Matthew, and Devin Frank. 2018. "Propositions." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2018/entries/propositions/

Moltmann, Friederike. 2013. *Abstract Objects and the Semantics of Natural Language*. Oxford: Oxford University Press.

Nelson, John O. 1983. "Do Animals Propositionally Know? Do They Propositionally Believe?" *American Philosophical Quarterly* 20 (2): 149–60.

Pickel, Bryan. 2017. "Are Propositions Essentially Representational?" *Pacific Philosophical Quarterly* 98 (3): 470–89. https://doi.org/10.1111/papq.12123.

Rosen, Gideon. 2020. "Abstract Objects." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/abstract-objects/.

Sacchi, Elisabetta. 2006. "Fregean Propositions and Their Graspability." *Grazer Philosophische Studien* 72 (1): 73–94. https://doi.org/10.1163/18756735-072001004.

Schiffer, Stephen. 2012. "Propositions, What Are They Good For?" In *Prospects for Meaning (Current Issues in Theoretical Philosophy, Vol. 3),* edited by Richard Schantz. Boston, Mass.: Walter de Gruyter.

Soames, Scott. 2010. *What Is Meaning?* Princeton, NJ: Princeton University Press.

Suddendorf, Thomas, and Michael C. Corballis. 2007. "The Evolution of Foresight: What Is Mental Time Travel, and Is It Unique to Humans?" *Behavioral and Brain Sciences* 30 (3): 299–313. https://doi.org/10.1017/S0140525X07001975.

Textor, Mark. 2017. "Judgement, Perception and Predication." In *Act-Based Conceptions of Propositional Content: Contemporary and Historical Perspectives*, edited by Friederike Moltmann and Mark Textor. Oxford: Oxford University Press.

Wetzel, Linda. 2018. "Types and Tokens." In *The Stanford Encyclopedia of Philosophy,* edited by Edward N. Zalta, Fall 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entriesypes-tokens/

Wierzbicka, Anna. 1996. *Semantics: Primes and Universals*. Oxford: Oxford University Press.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Oxford: Blackwell.

Zuberbühler, Klaus. 2018. "Combinatorial Capacities in Primates." *Current Opinion in Behavioral Sciences, The Evolution of Language*, 21 (June): 161–69. https://doi.org/10.1016/j.cobeha.2018.03.015

Organon F 29 (1) 2022: 78–108
https://doi.org/10.31577/orgf.2022.29104

ISSN 2585-7150 (online)
ISSN 1335-0668 (print)

RESEARCH ARTICLE

Robert Kirk's Attempted Intellectual Filicide: Are Phenomenal Zombies Hurt?

Dmytro Sepetyi*

Received: 6 June 2020 / Revised: March 30 / Accepted: 24 May 2021

Abstract: In the paper, I discuss Robert Kirk's attempt to refute the zombie argument against materialism by demonstrating, "in a way that is intuitively appealing as well as cogent", that the idea of phenomenal zombies involves incoherence. Kirk's argues that if one admits that a world of zombies z is conceivable, one should also admit the conceivability of a certain transformation from such a world to a world z* that satisfies a description D, and it is arguable that D is incoherent. From which, Kirk suggests, it follows that the idea of zombies is incoherent. I argue that Kirk's argument has several minor deficiencies and two major flaws. First, he takes for granted that cognitive mental states are physical (cognitive physicalism), although a zombist is free to—and would better—reject this view. Second, he confuses elements of different scenarios of transformation, none of which results in the incoherent description D.

Keywords: Consciousness; conceivability; incoherence; materialism; phenomenal zombie; possibility.

* Zaporizhzhia State Medical University
https://orcid.org/0000-0003-2110-3044
Maiakovskyi avenue, 26, Zaporizhzhia, Ukraine, 69035
dmitry.sepety@gmail.com

1. Introduction

In 1974, Robert Kirk introduced the concept of a phenomenal zombie—a creature physically exactly just like a conscious human being but without subjective experiences—and used this concept in an argument against physicalism. The argument, in brief, is that because zombies are conceptually or logically possible, phenomenal consciousness is something more than anything merely physical, so physicalism is false. When Kirk formulated the argument, it did not draw much attention. However, the argument was revived and made famous by David Chalmers in his philosophical bestseller *The Conscious Mind* (1996) and later papers. In the meantime, Kirk reversed his views and joined anti-zombists. In a series of publications, he argued that zombies are impossible in the relevant sense. He made the fullest exposition and defense of his argument in the book *Zombies and Consciousness* (2005), and recapitulated the argument in the paper "The inconceivability of zombies" (2008), and most recently in the book *Robots, Zombies and Us* (2017).

In *Zombies and Consciousness*, Kirk writes that one of the two main aims of the book is nothing less than "to dispose of the zombie idea once and for all". Not that there were no attempts to undermine the idea of zombies before. Kirk notes that "there are plenty objections to it in the literature, but they lack intuitive appeal". He believes his own attack on zombies fares better: "I have an argument which I think demolishes it [the zombie idea] in a way that is intuitively appealing as well as cogent" (Kirk 2005, vii).

In this article, I analyze Kirk's argument and show that it falls far short of the target. I will first outline the argument as presented in (Kirk 2005) and make some relevant elucidations. I will then note a few ways in which the formal representation of the argument needs clarification and more precise formulation (in accordance with Kirk's own explanations). Next, I will clarify the relationship between the idea of phenomenal zombies and interactionism. From this, I will proceed to the exposition of the failure of Kirk's argument. Finally, I will explain, in brief, why Kirk's later expositions of his argument (Kirk 2008; Kirk 2017) fare no better.

2. Kirk's argument in outline

As a preparation to his attack on zombies, Kirk argues, together with the supporters of the zombie argument (further on to be called "zombists"), that the possibility of zombies, in the sense that the idea of zombies involves no "inconsistency or other incoherence of a broadly logical or conceptual kind" (Kirk 2005, 10), is inconsistent with physicalism (materialism): if zombies are possible in this sense, then physicalism (materialism) is false.

Kirk (2005) calls that kind of possibility "c-possibility"; Chalmers (1996; 1999; 2004) and later Kirk (2017, 75-92) call it "logical possibility". It should be distinguished from the possibility in a much more limiting sense, "natural" or "nomical" possibility—what is possible *given the laws of nature operant in the actual world*. Zombies may be naturally impossible (the laws of nature operant in the actual world ensure that whenever there are some physical states, there are some mental states) but c-possible, in the sense that there is no incoherence in the idea of zombies (or of a world inhabited by zombies instead of conscious human beings); the physical facts do not, on their own, entail there being consciousness (phenomenal mental states). If that is so, then consciousness is something extra, besides the physical.

If Kirk's arguments to this point are sound (and I think they are), then the c-possibility of zombies entails that materialism (physicalism) is false. If so, then to defeat the zombie argument, materialists should defeat the claim that zombies are c-possible, that is, that the idea of zombies is coherent. Because the claim has a strong intuitive appeal, the physicalist is invited to present the case to the contrary that is at least as (or, better yet, more) strongly appealing. Can she do this?

One helpful remedy Kirk recommends against the zombie idea is noting that it (at least, in its initial pure form) conflicts with another intuition that is at least as strong, or even stronger—the view that consciousness matters for our behavior. It seems absurd to think that our behavior, including what we talk and write, does not depend on our consciousness, so that consciousness could be subtracted and this would make no difference for what we do, including what we say and write. How could zombies, for example, talk about their phenomenal mental states, *qualia*, feels, if they had none? However, (this is not Kirk's point but mine) for a person who

would find the zombie argument persuasive apart from this conflict, the conflict seems to play for interactionist dualism rather than for materialism. On the one hand, the zombie possibility intuition (basically, the intuition that nothing physical entails anything subjective) rules out physicalism; on the other hand, the consciousness-matters-for-behavior intuition rules out epiphenomenalism; and both intuitions, if correctly understood, agree with interactionist dualism.

Kirk is aware that the zombie argument is not an argument for epiphenomenalism (to exclude all other possibilities) but an argument against physicalism (materialism) that leaves open the choice between epiphenomenalism, interactionism, panpsychism, and idealism. At least, it is so construed by the most eminent contemporary zombist, David Chalmers, who explained that the conclusion of the argument "is the disjunction of panprotopsychism, epiphenomenalism, and interactionism" (Chalmers 1999, 493). Kirk takes this into account. His argument is intended to be equally demolishing against all sorts of zombists, be they epiphenomenalists or interactionists or whoever.

The target of the argument is the claim that zombies are conceivable. If successful, the argument shows that although zombies seem conceivable, they are really inconceivable. It should be noted that the word "conceivability" is a red herring here; all that really matters is c-possibility, that is, the coherence of the idea. "Conceivability" pops up because David Chalmers construed the zombie argument as involving the subservient argument: (1) zombies are conceivable; (2) conceivability entails possibility; therefore, zombies are possible. Commenting on this argument, Kirk remarks that both of its premises are obscure because it is not clear how to understand "conceivable"; no clear meaning was ever specified in such a way that both premises were difficult to challenge. The general situation is that "the lower the threshold for conceivability, the easier it is to accept premiss (1)—but the harder it is to accept premiss (2)". Whatever the case, "to prove c-impossibility ... must be a good way to prove inconceivability" (Kirk 2005, 27). So, Kirk proceeds to prove that zombies are c-impossible in order to prove that they are inconceivable. However, why did he need to prove that zombies are inconceivable? Obviously, in order to block Chalmers' argument that since zombies are conceivable, they are c-possible. However, he

need not do it at all, if *he has already proven that zombies are c-impossible* (≡the notion of a zombie is incoherent). So, let us put aside "conceivability", and in further discussion, whenever Kirk uses the word, replace it with "c-possibility" or simply "possibility", or stipulate that "conceivability" means the same.[1]

Kirk's purported proof of the c-impossibility of zombies involves a story he calls "e-qualia story" and two claims about it:

(C1) the c-possibility of zombies entails the coherence of the e-qualia story,
(C2) the e-qualia story is incoherent.

From (C1) and (C2), it follows that that zombies are c-impossible (there is a hidden contradiction in the idea of zombies). The argument is valid. However, is it sound? Are both (C1) and (C2) true? Kirk meticulously argues that they are. I am going to agree, by and large (with some qualifications), with (C2) and show that Kirk's argument for (C1) fails.

The e-qualia story is a story of a (possible, or perhaps impossible) world that satisfies the following conditions:

(E1) The world is partly physical, and its whole physical component is closed under causation: every physical effect has a physical cause. ...
(E2) Human beings stand in some relation to a special kind of non-physical properties, e-qualia. E-qualia make it the case that human beings are phenomenally conscious.
(E3) E-qualia are caused by physical processes but have no physical effects: they could be stripped off without disturbing the physical world.
(E4) Human beings consist of nothing but functioning bodies and their related e-qualia.

[1] In recent personal email communication, Chalmers informed me that he now defines conceivability as not-apriori-not and has abandoned the use of the term "logical possibility". If so, "conceivability", in Chalmers' present use, is the same as Kirk's (2005) c-possibility and Kirk's (2017) logical possibility. (The treatment of the issue in (Chalmers 2002) and (Chalmers 2010) can be best interpreted along these lines.)

(E5) Human beings are able to notice, attend to, think about, and compare their e-qualia. (Kirk 2005, 40)

A small terminological correction seems to be appropriate here. It would be more correct to talk in the e-qualia story not about "human beings" but, like in the zombie-story, about "the human-like inhabitants" of the e-qualia world, because if human beings in the real world do not satisfy the conditions (E2)-(E5), those e-qualia world inhabitants that satisfy them would perhaps not qualify as "human beings". However, this disqualification does not affect the argument. We may introduce the name "hubes" to designate both human beings and human-like inhabitants of possible (or even impossible) worlds that are exactly (or as far as possible exactly) like human beings physically, although they may essentially differ from human beings in other respects, which have to do with something non-physical (such as e-qualia). Now, having the term, let us replace in the e-qualia story "human beings" with "hubes".

To have convenient reminders for later use, let us designate the clauses (E3), (E4), and (E5) as "INERTNESS", "HUBES' COMPOSITION", and "EPISTEMIC CONTACT".

Kirk first argues for the claim (C2), that the e-qualia story is incoherent, and then for (C1), that if zombies are c-possible, then the e-qualia story should be coherent; if this argumentation succeeds, it follows that zombies are c-impossible. The argument for (C1) begins with a zombie story—a description (which should be coherent if zombies are c-possible) of a zombie world *z* that satisfies the following conditions:

(A1) z is purely physical, causally closed system;
(A2) Physically, *z* is as far as possible exactly like the actual world;
(A3) The human-like inhabitants of *z* lack phenomenal consciousness. (Kirk 2005, 49)

To have a convenient reminder for later use, let us designate (A2) as "PHYSICAL IDENTITY".

Kirk proceeds to argue that *z* can be transformed—in a way that zombists should acknowledge to be c-possible, namely, by adding to it the non-physical factor that is supposedly responsible for consciousness in the actual world or, at least, one that is the same insofar as the phenomenology of

non-physical mental states (e-qualia) is concerned[2]—into a z^* world that satisfies five conditions (Z1)-(Z5) that are equivalent to the conditions (E1)-(E5) of the e-qualia story. However, the preceding argument has established that the e-qualia story is incoherent; therefore z*-story is incoherent. However, z*-story is derived from the zombie-story and (if Kirk's arguments to the point are sound) a coherent transformation-story, from which it follows that the zombie-story is incoherent. Therefore, zombies are c-impossible, *q.e.d.*

Let us consider Kirk's argument in more details.

3. The e-qualia story: clarifications and reformulations

3.1. "E-qualia", qualia, and cognitive mental states

Apparently, the term "e-qualia" implies *qualia*—specific subjective qualities of mental states, their "what-it-is-likeness" for a mental subject (experiencer); the prefix "e-" probably is a shorthand for "epiphenomenal". Usually, when talking of *qualia*, philosophers mean *subjective qualities of experiences*, such as painfulness of pain (how it feels), or what it is like for an experiencer to have an experience of red color, or some other experience. Quite a few philosophers think that only qualia in this limited, experiences-bound sense are strongly resistant to materialistic (physicalist, or functionalist) reduction, whereas other aspects of mind, having to do with cognition and meaning, are far less problematic; therefore, we can assume that cognitive capacities (such as to notice, attend to, think about, and compare) can be fully accounted for in materialist (functionalist) terms, and focus on experiential qualia. (Further on, let us refer to this view as "cognitive physicalism".)[3] However, that is exactly what we *should not do* when discussing

[2] In Kirk's own description, it should be "a non-physical item or items x which, when appropriately associated with z, would ensure that its inhabitants acquired our kind of phenomenal consciousness" (2005, 49).

[3] See, for example, (Levine 2001, 4-6). In (Chalmers 1996), cognitive physicalism is not stated explicitly but seems to be presumed implicitly, in his distinction of psychological and phenomenal properties (where "psychological properties" are defi-

arguments against materialism, because this approach limits our choice to materialism and its most emasculated alternatives, and keeps out of discussion the more robust and defensible alternatives. To take this approach means to give materialism the uncontested lordship over the largest and most important part of the mental realm, and leave it for non-materialists to fight for the remaining poor grounds. Such a fight would be nearly hopeless, for in it, materialists would have on their side all the advantages, and their opponents all the disadvantages, of assuming that there is nothing to cognitive mental states besides physical processes that fulfill certain functions.

I think that the remark made by Howard Robinson in a paper defending the knowledge argument is relevant for the zombie argument as well:

> Those who ... think that physicalism can be correct for everything but qualia are in inconsistent position. The knowledge argument should not be cast in the form "physicalism can work for all other mental states but not for qualia", but in the form "even if it might look as if functionalism will work for less clearly introspectible states, such as thoughts, Mary's case shows that it will not work for qualia, and we can see from this that it does not work for thought—at least, a certain category of thought...—either." (Robinson 2004, 72)

On the most natural, common-sense and strongly intuitively appealing view, our (conscious) thinking, understanding, and willing are intrinsically

ned purely functionally, as ones that "play the right sort of causal role in the production of behaviour" (p. 11)), with putting awareness and other cognitive states on the "psychological" side, and especially in the thought experiments of fading and dancing qualia (pp. 253-274). However a year later, in reply to the criticisms advanced by Hodgson, Lowe, Velmans, and Libet, Chalmers repudiated cognitive physicalism and proposed that the use of cognitive terms in his earlier writings should be taken "in a stipulative sense" rather than as assuming that there is nothing more to cognitive states besides their behavioral functionality (Chalmers 1997, 20). However, I think that such a reading leaves his thought experiments of fading and dancing qualia deficient. It also may be relevant to note that (Chalmers 1997) and later is far more favorable to interactionism than (Chalmers 1996), and that the arguments of fading and dancing qualia presume epiphenomenalism.

just as subjective as an experience of pain or of green color. The idea that all those physical processes that go on in human bodies and brains could (in the sense of c-possibility) occur without there being any subjective (conscious) awareness and understanding is just as plausible as the idea that C-fiber firing in the brain could occur without pain-sensation. The (*prima facie* coherent) concept of the phenomenal zombies implies that the zombies lack not only the capacities for such subjective experiences as pain-qualia or red-color-qualia but also the capacities to notice, attend to, think about, and compare in any (human) sense that involves subjective (conscious) awareness and understanding. At the very least, such a view is open (and, I suggest, commendable) for a zombist. A zombist would do well to posit *subjective* (conscious) cognitive states (processes) of thinking-awareness-understanding on the phenomenal, not the physical side. If zombies are c-possible, then the states of "noticing, attending to, thinking about, and comparing", in the sense relevant to the zombie argument, belong to the category of "a special kind of non-physical properties" that "make it the case that human beings are phenomenally conscious", that is, "e-qualia", on Kirk's definition (E2).

3.2. Where does the incoherence lie? Direct and indirect causation

Kirk argues that (E3), INERTNESS, is inconsistent with (E5), EPISTEMIC CONTACT. In fact, his argument goes through only if the beginning clause of (E3), "E-qualia are caused by physical processes", is understood as "*All* e-qualia are *directly* caused by physical processes *alone*". What Kirk really argues for is that if e-qualia have no effects whatever, whether physical or non-physical, then it is impossible for hubes to be able to notice, attend to, think about, and compare their e-qualia. There is no need to delve into details of Kirk's argument to this point. For my purposes, it is enough that the claim is *prima facie* very plausible: how can I attend to my experiences (assumed to be non-physical e-qualia), or think about my experiences, if my experiences never cause, or play any causal role in causing my attention or thinking?

On the other hand, if we take (E3) literally, without the qualifications "all", "directly", "alone", then Kirk fails to make the case that the e-qualia

story is incoherent. Kirk's argument for the incoherence of the e-qualia story (the inconsistency between INERTNESS and EPISTEMIC CONTACT) rests entirely on the absence of causal connection from experiences to attention and thinking. However, Kirk's formulation of the story (especially, with respect to INERTNESS) does not exclude the possibility that hubes' non-physical experiences are causally relevant to their attention and thinking, *if attention and thinking are taken to be non-physical* mental states (even if they are epiphenomenal, having no physical effects).

So, the dilemma arises:

either Kirk's argument fails to show that the e-qualia story is incoherent,

or the formulation of the e-qualia story should be made more precise so as to exclude *any* possibility of there being a causal link from experiences to attention and thinking, whether the latter are taken to be physical or non-physical.

If the former, then the argument fails full stop. If the latter, the argument can proceed with the e-story slightly reformulated, by inserting into (E3) the qualifiers "all", "directly", "alone":

(E3*) INERTNESS* *All* e-qualia are *directly* caused by physical processes *alone* but have no physical effects: they could be stripped off without disturbing the physical world.

It may be objected here that in fact, Kirk (2005, 42) *does argue* that the e-qualia story forbids e-qualia to have effect on other e-qualia. However, if you consider the argument, you easily see that it is made on the construal of (E3) as (E3*). The argument is that "since by (E3) all qualia are already caused to occur by physical events", there would be no work for e-qualia-to-e-qualia causation to do. Now three points should be noted.

1) In fact, the qualifier "all" is absent in the formulation of (E3) on p. 40, but it is clear that because there is no other qualifier (such as "some"), "all" is implied. And on p. 42 Kirk confirms this explicitly. So my adding "all" does not really change (E3) but merely emphasizes its point.

2) There is no causal work for e-qualia to do only if all e-qualia are caused by physical processes *alone*, in the sense that physical processes *are*

sufficient for causing these e-qualia—to be distinguished from the possibility that some e-qualia are produced jointly by physical processes and e-qualia, so that without the participation of e-qualia physical processes would not have this effect. So, adding the qualifier "alone" (understood in this sense) is perfectly justified.

3) There is no causal work for e-qualia to do only if all e-qualia are caused by physical processes *directly*,—to be distinguished from indirect causation, when a physical process P causes an e-qualia A that, in its turn, causes an e-qualia B. If P causes A that causes B, then the fact that P causes B (by causing A that causes B) in no way robs A of its causal work. So, if anything, the argument on p. 42 shows that *in Kirk's own meaning, (E3) should be construed as (E3*).*

3.3. Robinson's objection

Howard Robinson (2016, 55) proposed that a zombist can deny the incoherence of the e-qualia story by 1) making use of the typical functionalist account of intentionality, according to which the intentionality of an epistemic state is a matter of behavioral appropriateness, and 2) holding that this behavioral appropriateness need not necessarily be due to an epistemic state's being caused by its object (experience, in our case) but can be due to common causal ancestry of both the epistemic state and its object. I think Robinson is right that Kirk does not provide an argument to neutralize such a move. However, for me, personally, it seems highly plausible that for an epistemic state to be about a particular real object (at least, in the sense of original intentionality), there must be causal link from the latter to the former. So although the move proposed by Robinson is available for a zombist, I propose to explore the availability of other resources.

For convenience of the following discussion, it is useful to introduce a distinction between three possible varieties of "zombists" that would treat Kirk's argument differently. Let us designate a zombist who is *not a cognitive physicalist* "a Cartesian zombist" (because he/she, like Descartes, holds that thinking pertains to a non-physical mind rather than to a physical body), and a zombist who is a cognitive physicalist—"a non-Cartesian zombist". With non-Cartesian zombists, the way to meet Kirk's argument depends on whether a zombist is an epiphenomenalist or an interactionist. A

non-Cartesian epiphenomenalist in fact holds that the actual world satisfies the e-qualia story, and so he/she can meet Kirk's argument only by denying that the e-qualia story is incoherent (probably, in the way Robinson suggests). With respect to such a zombist, the rest of Kirk's argument has nothing to do. So the following discussion is concerned only with the coherence of Cartesian zombism and non-Cartesian interactionist zombism.

4. The zombie story and interactionism

In the zombie story (A1)-(A3), the condition (A2), PHYSICAL IDENTITY, stipulates that the zombie world to be discussed (z) is physically "as far as possible exactly like the actual world". The phrase "as far as possible" needs an explanation: why did Kirk moderate his zombie story with it? The purpose was to take into account Chalmers' explanation that the zombie argument is not an argument for epiphenomenalism but leaves open several non-materialistic alternatives, such as panprotopsychism, epiphenomenalism, and interactionism. If so, for Kirk's argument to bite against zombism generally (not only against epiphenomenalist zombism), the description of a zombie world *z* should be such that any zombist (interactionist as well as epiphenomenalist) should admit the c-possibility of such a world.

How can non-epiphenomenalist views be reconciled with the possibility of zombies? In particular, how can it be with interactionist dualism? *Prima facie*, it seems that interactionism is inconsistent with the c-possibility of zombies. It seems that if in the actual world, non-physical consciousness causally influences brain states responsible for behavior (as interactionists believe), phenomenal zombies and zombie-worlds as usually described are c-impossible for an obvious reason: zombies lack some *physically relevant* (although non-physical) causal factor that we have, and so the physical dynamics of their bodies' functioning should be different. However, David Chalmers explained how "the possibility of zombies is compatible with non-epiphenomenalist dualism": "an interactionist dualist can accept the possibility of zombies, by accepting the possibility of physically identical worlds in which physical causal gaps (those filled in the actual world by mental processes) go unfilled, or are filled by something other than mental processes" (Chalmers 2004, 182-183).

4.1. Replaceabilism

However, "the possibility of physically identical worlds in which physical causal gaps (those filled in the actual world by mental processes) go unfilled, or are filled by something other than mental processes" is likely to seem problematic, at least *prima facie*. One may think that the idea of a world in which some physically relevant causes are systematically lacking but all physical events go as if nothing were lacking is incoherent; such a world is not c-possible. Otherwise, if in a possible world, physical causal gaps are filled by something other than mental processes, what can this "something other" be? It seems that if it is not mental and is causally relevant, there is no reason why it should not count as physical. However, if it counts as physical, then the possible world so conceived is not exactly physically identical with the actual world; it has some physical surplus. On the other hand, we can run the zombie argument with a modification that takes care of such a physical surplus: if zombies with a physical surplus are c-possible, it seems that materialism should be false, because those zombies lack nothing physical that human beings have but lack consciousness. (It is implausible that adding some physical surplus would bereave human beings of consciousness and turn them into zombies). Although Kirk did not go in these details, he made his description of the zombie world z in such a way that it could accommodate such zombies with a physical surplus (and so make it possible for some interactionists to count as zombists), by means of the phrase "as far as possible" in the condition (A2).

An interactionist zombist view that is so accommodated can be designated as *replaceabilism.* A replaceabilist admits the possibility of zombies with a moderate modification to the initial (Kirk 1974a; Kirk 1974b; Chalmers 1996) specification. Replaceabilist interactionism is consistent with, and implies, the c-possibility of *modified* phenomenal zombies or a *modified* zombie world that *lacks nothing physical* that we (or the actual world) have but lacks consciousness nevertheless. There just should be, in those modified zombie worlds, some other *physically relevant* causal factors to compensate for the causal deficiency resultant from the subtraction of human consciousness. By the condition (A1), which says that a zombie world z is purely physical, Kirk stipulates that these factors should themselves be physical (so, a zombie world may be physically richer than the actual world).

As an alternative to filling the causal gap (that should—if interactionism is true—result from subtracting mental processes) with some additional physical factors, we can conceive of some possibility like the following. Imagine a world that runs in parallel with ours and is at every moment exactly like our world a minute ago in all physical respects, because this world is governed by a physically omniscient and omnipotent demon who took fancy to support that belated-copy-world so that all physical deviations (that may happen because the humanlike inhabitants of that world have no consciousness, or because of quantum-mechanic indeterminacy) are almost instantly detected and eliminated by the demon. Although in such a conceivable scenario, some mental processes (those of human beings, indirectly, and those of the demon, directly) are causally efficient with respect to physical events in the zombie world, the zombies themselves are purely physical copies of human beings without phenomenal mental states, so they fit the requirements of the zombie argument.

Besides replaceabilism, an interactionist dualist has two other options, which I designate as *irreplaceabilism* and *supercoincidentalism*. Let us consider these alternatives and their relationship with the c-possibility of zombies.

4.2. Irreplaceabilism and the conditional construal of the zombie argument

An interactionist can deny the c-possibility of replacing human phenomenal minds with some physical entities so that all physical events proceed without any change.[4] Kirk mentions such a possibility and remarks that

[4] It is open for an interactionist—at least, if he is a substance dualist—to take the view that the human mind, or self, or soul develops and affects the brain in such a way that it is in principle (as a matter of c-possibility) irreplaceable—not with respect to some particular effect but with respect to *all the totality* of its *real and possible* physical effects *throughout the life*—with anything physical.

Is irreplaceabilism plausible? I think that it is. To see this, let us first think of our talks, and writings, and philosophical discussions about our experiences and other conscious states and processes (such as having a certain occurrent thought). It seems very implausible that all the physical aspect of all these happenings could be

"some interactionists might deny that physical events could cause humanlike behavior, but they could not be zombists" (Kirk 2008, 85). This should be admitted, given Kirk's definition of "zombists" as "those who think zombies are conceivable" (Kirk 2005, 38), where "conceivable"="c-possible". However, such an interactionist—zombist or not—can still find use for the zombie idea and the zombie argument in a conditional way, as suggested by Andrew Bailey, "as part of a destructive dilemma for the physicalist": either physical reality is *not* causally closed, and so physicalism is false, or it is causally closed, and then zombies are possible, and so physicalism is false anyway (Bailey 2009, 135).[5] If so, then Kirk's argument falls short of his most ambitious purpose of "disposing of the zombie idea once and for all", or demolishing it (Kirk 2005, vii), even if it were successful in all other respects (which it is not, as will be shown in the following sections), that is, against all those who fall under his definition of "zombists".

effected by zombies without any experiences and other conscious states and processes, with some purely physical substitute. It is far from clear (and, I think, implausible) that a purely physical substitute for consciousness capable of such an achievement is possible, even in principle (as a matter of c-possibility). And this becomes even more so, if we think of such persons as Plato, or Einstein, and their intellectual achievements, and the impacts of those achievements on the course of human history, behaviors of millions of people, etc. Presumably, their intellectual achievements were a matter of conscious interest to some problems, conscious understanding, conscious thinking, and conscious guess. Presumably, their huge impact on the human history, on behaviors of millions of people, was a matter of other people's conscious interests and understanding, etc. Is it possible, even in principle, (c-possible) that in a modified zombie-world, its humanlike inhabitants-zombies would do all the same movements, with all the same (speechlike) sounds produced, books written and typed, machines and computers produced and run, as a result of nothing but purely physical interactions of the microphysical components of which their bodies consist, with no (phenomenal) consciousness at all? Perhaps it is, but it is at least just as plausible that it is not.

[5] As Kirk himself remarks, "the idea of zombies suggests itself as soon as one accepts the causal closure of the physical" (Kirk 2008, 74). Thus, an irreplaceabilist can consider the zombie argument as showing not what *is* logically possible (given that the world is such as it is, that is, interactionistic) but what *should be* logically possible on the assumption that the actual world is causally closed with respect to the physical events.

4.3. Supercoincidentalism

Alternatively, the interactionist zombist can hold that even a non-modified zombie world (with no physical entities added and no non-physical factors involved) is possible, but such a possibility involves a superhugely improbable succession of coincidences (infinitely more improbable than the chance that a tornado sweeping through a junkyard might assemble a Boeing 747). The point is that if the idea of a genuine *physical* causal indeterminacy is not incoherent (and quantum mechanics seems to show that it is not merely coherent but holds in the actual world), and if consciousness has physical effects in the actual world, then it is not strictly impossible that there may be an exact physical duplicate of the actual world in which there is no consciousness: in that world, all physical events turn out to be the same as in the actual world as a result of a superhugely improbable—but not strictly impossible—quantum-mechanical flukes.

Take note: it is not the case that quantum mechanical indeterminacy applies only to microphysical but not to macrophysical events. It applies throughout the board, only that for macrophysical events, the probability of a considerable deviation from the "normal" deterministic course is hugely small. A zombie world is a world in which such hugely improbable events regularly happen with zombies, and incidentally they happen in such a way that all parts and particles of zombies make exactly the same movements as the corresponding parts and particles of human bodies in the actual world.

Supercoincidentalism has a considerable advantage over the other two interactionist options, in that (1) it accommodates the c-possibility of zombies in the most direct way (which requires no modification to the initial specification of zombies) and (2) it saves the irreplaceabilist intuition that it is superhugely unlikely that purely physical twins of human beings with no phenomenal minds would behave in all exactly the same ways as conscious human beings do throughout whole human lives.[6]

[6] Consider the infinite set of possible worlds that at some moment ***t*** are exact physical copies of the actual world at this moment but in which there are no phenomenal minds. In that set, the subset of worlds in which all physical events with zombies will for a considerable time proceed exactly as they do in the actual world

Although the supercoincidentalist option is distinct from the replaceabilist one, the argument that follows fits both in the same way.

5. The transformation story, and where Kirk's argument fails

Zombists hold that a zombie world z described by the conditions (A1)-(A3) is c-possible. Should a zombist agree with Kirk that adding to z the non-physical factor (from now on to be called "the consciousness factor") that is supposedly responsible for consciousness in our world, or its "phenomenal duplicate" (perhaps, bereaved of powers to produce physical effects), can conceivably transform that world into the world z^* identical to the e-qualia story world?

Consider Kirk's description of z*:

(Z1) z^* is partly physical, and its whole physical component is closed under causation: every physical effect in z^* has a physical cause.

(Z2) The human-like organisms in z^* are related to a special kind of non-physical item x. x makes it the case that they are phenomenally conscious.

(Z3) x is caused by physical processes but has no physical effects: it could be stripped off without disturbing the physical component of z^*.

(Z4) The human-like inhabitants of z* consist of nothing but functioning bodies and their related x. […]

(Z5) The human-like inhabitants of z* are able to notice, attend to, think about, and compare the qualities of their experiences. (Kirk 2005, 51)

In the description, "x" stands for the consciousness factor.

For our discussion, there are two important questions about this description to be asked and answered:

with human beings makes up an infinitely small—going to zero—portion. The probability of hitting at random at such a world in such a set goes to zero. Nevertheless such zombie worlds are not strictly impossible.

- Why does Kirk think that a zombist is committed to the c-possibility of the transformation from z to z^*?
- Is (Z1)-(Z5) really equivalent to the e-qualia story?

5.1. The epistemic intimacy argument and its failure

Why does Kirk think that a zombist is committed to the c-possibility of the transformation from z to z^*? In brief, his reasons are as follows.

If one admits that z is possible, then one cannot deny that z^+ is possible, where z^+=z+the consciousness factor, where the consciousness factor is phenomenally just like human consciousness, has the same dependence on human brains, and *has no physical effects*. This ensures that in z^+, (Z1), (Z2), (Z3), (Z4) hold (Kirk 2005, 49-50).

Kirk argues that zombists should admit the c-possibility that (Z5) holds as well: because z^+ is exactly like the actual world *physically*, and it has the consciousness factor x that is phenomenally exactly like "the non-physical item y which they think produces phenomenal consciousness in the actual world", and given that we have *epistemic intimacy* with our experiences, there is nothing to account why the hubes of z^+ cannot c-possibly have such epistemic intimacy with their experiences (Kirk 2005, 50-51). I suggest that Kirk could strengthen his argument by pointing out that because z^+ is physically identical with the actual world, its hubes would talk and write about their experiences just like we do, and this is impossible if they are unable to notice, attend to, think about their experiences. Let us designate this argument as *the epistemic intimacy argument*.

If *the epistemic intimacy argument* succeeds, then a zombist should admit the c-possibility of z^*.

It should be noted that Kirk's argument is made *on the assumption of cognitive physicalism*: it is stipulated that the consciousness factor in z^* has no causal impact on the physical but is not stipulated that there are no causal connections *within* the consciousness factor. A bit later, I will argue that a zombist who is a cognitive physicalist (a non-Cartesian interactionist zombist) can plausibly decline the epistemic intimacy argument, and so Kirk's purported refutation of zombism fails. However, I will first explore how a Cartesian zombist (who is *not a cognitive physicalist*) can respond Kirk's argument.

A Cartesian zombist would not need to resist the epistemic intimacy argument, insofar as the latter stipulates that the consciousness factor x in z^+ has no *physical* effects and does not stipulate that there is no causation within x itself. Insofar as such intrinsic causation within the consciousness factor x is not ruled out, a Cartesian zombist would agree that a world z^* that satisfies (Z1)-(Z5) is c-possible. You just add something like a Cartesian soul (having cognitive mental states as well as experiences) to z and deprive it of all powers to produce physical effects. The resulting world would satisfy (Z1)-(Z5); however, there is no incoherence involved. This is possible because (Z1)-(Z5), although very much like the e-qualia story, is not exactly the same. (Z3) is not quite the same as (E3), and (Z4) is not quite the same as (E4). However, the dialectics of the argument will come to the point where a Cartesian zombist can be confronted with *the modified epistemic intimacy argument* (and the modified clause ($Z3^m$)) that involves the stipulation that nothing in the consciousness factor x has *any* effects, *physical or nonphysical*. (Note that ($Z3^m$) indeed would be equivalent to (E3*), INERTNESS* in the e-qualia story.)

So, a zombist who admits the incoherence of the e-qualia story should find something wrong with the *epistemic intimacy argument*, either initial or modified or both.

Happily for a zombist, there is a simple explanation as to what is wrong with these intimacy arguments. It is as follows.

Although there are no behavioral (and generally physical) and no phenomenal differences between z^+ and the actual world, there still can be *some* differences that make it the case that (Z5) cannot hold in z^+. What other differences can there be, given that both the actual world and z^+ contain nothing but physical entities and consciousness? The answer is that there is an important *difference in causal relations*: presumably, there is causation from experiences to cognitive mental states in the actual world; on the other hand, Kirk's argument hangs on the stipulation that there is no such causation in z^*. A zombist can hold that such an absence is inconsistent with (Z5)—at least, if the e-qualia story is indeed incoherent.

Recollect that the e-qualia story was found incoherent exactly because arguably, in the absence of causal connection from experiences to cognitive mental states, there can be no cognitive mental states *about experiences*:

the very absence of such causal relations rules out the existence of cognitive mental states *with such aboutness*. Surprisingly, Kirk fails to see that this applies to the c-possible result of adding the consciousness factor x to the zombie world z: if there is no causation from experiences to cognitive mental states, then (Z5) does not hold.

One can wonder: how can that be if z^+ is exactly like the actual world both physically (in particular, in behavior of its hubes) and phenomenally? The answer is that although there will be something it z^+ that is physically and/or phenomenally exactly like cognitive mental states about experiences in the actual world, that something *would not qualify as* cognitive mental states *about experiences*, because—at least, in cases when the referent is a particular really existent object—*the appropriate causal relationship is constitutive of aboutness* (at least, partially).

Again, that is exactly why in the e-qualia story (E3*), INERTNESS* seems to conflict with (E5), EPISTEMIC CONTACT. At least, in Kirk's argument for the incoherence of the e-qualia story, there is nothing to show that (E3*), INERTNESS* is inconsistent with there being physical states (including all behavioral movements) that are *physically exactly like* what a cognitive physicalist can take for cognitive mental states about experiences, or with there being some e-qualia that are *phenomenally exactly like* what a cognitive non-physicalist would take for occurrent cognitive mental states about experiences.

In fact, it is not too difficult to see how there can be two mental states that *are phenomenally identical but differ in their aboutness*: one is about a particular really existing thing, whereas the other is not. Take, for example, seeing a table and hallucinating a table. They can (c-possibly) be phenomenally the same, but the former is about a particular really existing table, whereas the latter is not. And this would be the case even if there really is a table in place where the hallucination suggests but that table has nothing causally to do with the hallucination. The same applies to cognitive states about particular real experiences and their c-possible phenomenal twins which *fail to be about particular real experiences*. Think of the fantastic Swampman-style scenario in which my physical duplicate gets assembled out of atoms. Suppose that I had a toothache yesterday, and I can well recollect that experience. Surely Kirk, as a materialist, should admit that

because we (and our brains) are physically exactly the same, my duplicate can "recollect" having a toothache yesterday, and this his "recollection" can be physically and phenomenally exactly like my recollection. However, my recollection is genuinely about a particular real experience I had yesterday, whereas—as Kirk himself argued—my duplicate's "recollection" cannot be genuinely about that (or any other real) experience, because there is no causal link from the experience to his "recollection". And although my duplicate can behave (move) exactly like I do when I talk or write about my past experiences, making just the same noises and leaving just the same marks on paper, this his behavior will not be talk and writing about his past experiences.

This crucial point made, in the rest of this section, I propose a detailed account of how the defense of a Cartesian zombist would proceed *before it arrives at the point of collision with the modified epistemic intimacy argument* (subsections 5.2-5.3), and an account of how a non-Cartesian interactionist should coherently envision the $z \rightarrow z^{+}$ transformation scenarios, none of which happen to result in z^{*} that satisfies (Z1)-(Z5) (subsection 5.4).

However, one of Kirk's latter expositions of his argument, (Kirk 2008), gives this discussion a new turn, to be discussed in section 6.

5.2. Broadening the e-qualia story

One objection that a zombist can make to Kirk's argument is that (Z4) and (E4), HUBES' COMPOSITION, are not equivalent. Compare

(Z4) Hubes of z^{*} consist of nothing but functioning bodies and their related consciousness factor x.

and

(E4) Hubes consist of nothing but functioning bodies and their related e-qualia.

A zombist is not committed to the view that the consciousness factor x is nothing but e-qualia—nor even to the c-possibility of such a consciousness factor. On Kirk's definition, e-qualia are non-physical properties; however, the description of z^{*} leaves it open that the conscious factor x may be more than that. Some zombists would prefer the substance dualism view that a

human beings consist of nothing but a functioning body and a *non-physical mental subject* that is a bearer of phenomenal mental states, or qualia. And they can hold that *it is not even conceivable for there to be e-qualia without nonphysical mental subjects that underlie them.* So, even if Kirk's claim that the c-possibility of a zombie entails the coherence of the description of z^* is right, this is not enough to refute zombism, because the description of z^* is not equivalent to the e-qualia story, at least insofar as (E4) and (Z4) are concerned.

However, Kirk anticipated this objection and dealt with it by claiming that the difference is not significant because "even if x were kind of unitary substrate rather than the collection of properties, it would have to underlie, realize, or otherwise provide for a plurality of properties", including phenomenal qualities, and "so far as the e-qualia story is concerned, therefore, those qualities might just as well be called 'e-qualia'" (Kirk 2005, 51-52).

The point of this reply is that the e-qualia story would still remain incoherent, if we supplement its ontology (which initially included only physical entities and e-qualia) with non-physical unitary substrates that (do nothing but) underlie e-qualia. Only in that case, the difference between (Z4) and (E4) is insignificant. (Surely, if adding such unitary substrates to the e-qualia story would make it coherent, then the difference would be very significant!)

If so, Kirk's reply amounts to the correction in the initial e-qualia story that can be made explicit by replacing (E4) with

(E4*) Hubes consist of nothing but functioning bodies and their related e-qualia, *or non-physical substances that are bearers of their related e-qualia.*

Given that e-qualia and a non-physical mental subject are the only candidates for the role of the consciousness factor x, with this modification to the e-qualia story (further on taken for granted), we have the required equivalence between (Z4) and (E4*). However, this does not save Kirk's argument.

5.3. Intra-mental causation and the defense of Cartesian zombism

There is a more important difference between (Z1)-(Z5) and the e-qualia story—the non-equivalence of (Z3) and (E3*), INERTNESS*. Compare:

(Z3) x is caused by physical processes but has no physical effects: it could be stripped off without disturbing the physical component of z^* (Kirk 2005, 51),

where x stands for the consciousness factor,
and

(E3*) All e-qualia are directly caused by physical processes alone but have no physical effects: they could be stripped off without disturbing the physical world.

(Z3) is *not equivalent* to (E3*) in an important respect. To see this, we should recollect that the e-qualia story was found incoherent because (E3*), INERTNESS* conflicts with (E5), EPISTEMIC CONTACT, and they were found inconsistent because INERTNESS* forbids non-physical mental events (such as pains or red-color qualia) to cause or take part in the causation of any cognitive mental states (such as conscious attention, thinking, etc.). However, a Cartesian zombist would find nothing in (Z3) to this effect. Nothing in the formulation of (Z3) forbids causation *within* the consciousness factor x. It is relevant to this point that the consciousness factor does not need to be a simple property (e-qualia). A zombist can hold (in line with a variety of property dualism) that the consciousness factor is a *set of causally connected non-physical mental states* (conscious experiences causing conscious awareness, attention, thought, etc.). Or she can hold that the consciousness factor is an entity-substance with rich internal differentiation, temporal development, and internal causal relationship between its states—it may be a full-blown mental subject, or self, or the Cartesian soul. If so, there is no incoherence between (Z3) and (Z5) analogous to the incoherence between (E3*) and (E5). So, the z^* story is not equivalent to the e-qualia story in a crucial way that makes Kirk's argument invalid.

Kirk was not ignorant of this kind of objection. He considered essentially the same objection (although formulated in different terms) and replied as follows:

> The question is not whether some metaphysical story or other could be told by which non-physical items were capable of cognitive processing. It is whether the conceivability of zombies entails the conceivability of the e-qualia story. Hence all I have to do is

> to show that if zombies are conceivable, then so is a version of (Z1)-(Z5) according to which *x* is inert. And this is a consequence of the fact that causation is a contingent matter. (Kirk 2005, 52)

Now, a zombist can object that this reply misses the mark. Kirk *did not show* that if zombies are c-possible, then so is the world *z*** describable by (Z1)-(Z5) *plus the additional condition* that the consciousness factor is inert in the strong sense required for his purpose—that besides having no external, physical effects, it also has no internal causal links between its own states, such as experiences and thinking about experiences.

Kirk did not foresee and answer this objection; however, at this point he could appeal to the epistemic intimacy argument appropriately modified (adapted to the context of cognitive non-physicalism). However, a Cartesian zombist can decline this argument in the way explained in the subsection 5.1. So Kirk fails to prove that Cartesian zombism is incoherent.

5.4. Causal overdetermination and the defense of non-Cartesian interactionist zombism

On the other hand, a non-Cartesian interactionist zombist (one who accepts cognitive physicalism) would consider two possibilities of envisaging the transformation from the zombie world *z* to the world *z**:

- either we add to *z* the same consciousness factor as that of the actual world, including its causal powers;
- or we add to *z* such a consciousness factor that is exactly like that of the actual world insofar as the production and phenomenology of its non-physical states (e-qualia) is concerned but is bereft of its causal powers to produce physical effects.

1) The case with the same consciousness factor, including its causal powers. In the first case, a zombist can point out that the resulting world cannot c-possibly fit (Z3), which says that the consciousness factor has no physical effects. The trouble with (Z3) is as follows.

In z*, physical factors alone have all the causal powers required to produce all the effects which the consciousness factor produces in the actual world. And in z*, the consciousness factor alone has all the causal powers required to produce all the effects it produces in the actual world. However,

in z*, causal powers of the physical are not alone, and causal powers of the consciousness factor are not alone; they are put together. It seems that this should result in different (additive) effects, as compared with those that would result if only one of them acted. (1+1 equals 2 rather than 1.) However, if the effects are different, then the consciousness factor is causally efficient, and (Z3) does not hold for z*.

However, Kirk could insist that it is conceivable (c-possible) that in z*, causal powers of the consciousness factor make no physical difference: the causal powers of the physical and the causal powers of the consciousness factor together produce exactly the same effect that each of them would produce alone. (In z*, causal 1 of the physical + causal 1 of consciousness equals causal 1, not causal 2.) A zombist can concede this, but point out that this would clearly be a case of causal overdetermination. If so, we still do not have (Z3); instead, we have, as the best c-possible approximation to (Z3)

(Z3*) x is caused by physical processes but has no non-overdetermined physical effects: it could be stripped off without disturbing the physical component of z*.

So, the envisaged world z* is not a world in which the consciousness factor has no physical effects but a world in which its physical effects are systematically overdetermined by physical factors. In this situation of overdetermined causation, the consciousness factor makes no physical difference; however, overdetermined causal links from the consciousness factor to physical states of the brain are still causal links, and there being such causal links may be a sufficient ground for some such brain states (or their functional aspect) to count as noticing, attending to, thinking about experiences, etc.[7]

2) The case with the consciousness factor's causal powers subtracted. In the second case, a zombist can point out that the resulting world does not

[7] A reminder may be appropriate that the interactionist at issue does not hold that there is such overdetermination in the actual world; he just holds that a world with such overdetermination is c-possible, and that in such a possible world, overdetermined causation from experiences to some physical brain states should count as sufficient for there to be *epistemic contact* from experiences to cognitive states.

fit (Z5), which says that the hubes "are able to notice, attend to, think about, and compare the qualities of their experiences". In the resulting world, there would be quasi-cognitive states that are physically (and so functionally) exactly like noticing, attending to, thinking about experiences, etc. in the actual world; however, they should not count as genuine noticing, attending to, thinking about experiences, etc., exactly because they do not stand in the appropriate causal relationship to experiences. (That is the case because the epistemic intimacy argument is mistaken, as was shown in the subsection 5.1.)

The remaining description (Z1)-(Z4), without (Z5), is crucially non-equivalent to the e-qualia story, because in the latter, the contradiction arises between (E1)-(E4*) on one side and (E5), EPISTEMIC CONTACT on the other. So the description of the world (Z1)-(Z4), unlike the (incoherent) e-qualia story, is perfectly coherent. And so Kirk fails to prove that non-Cartesian interactionist zombism is incoherent.

6. Kirk's later expositions of his argument. The necessity of epistemic contact with experiences: why Kirk would better not appeal to it

In his later paper, "The inconceivability of zombies" (2008), and again in the chapter 7 of his book *Robots, Zombies and Us* (2017), Kirk rehashes his argument in a bit different and less detailed way, with the same unquestioned implicit assumption of cognitive physicalism.

There are several differences to be pointed out.

1) Both (Kirk 2008) and (Kirk 2017) omit the epistemic intimacy argument altogether. They just take it for granted that adding to z an inert consciousness factor leaves the hubes of z* in epistemic contact with their experiences.

2) Nevertheless, (Kirk 2008) considers the possible objection on the side of an interactionist zombist that if the inert consciousness factor "continued to make our successors conscious, its lack of causal efficacy would prevent it from continuing to sustain epistemic contact" (Kirk 2008, p. 86), and makes a new argument against it.

3) In (Kirk 2017), the former e-qualia story goes under the name "epiphenomenalism".

Of these, only the second point can be taken as strengthening Kirk's position, so let us discuss it.

Kirk begins with the remark that he "find[s] it hard to make sense of that suggestion" (Kirk 2008, p. 86); then he quotes David Chalmers' statement that there is "not even a conceptual possibility" that a subject should have an experience "without any epistemic contact with it" (Chalmers 1996, p. 197), states his approval ("surely he is right about that") and adds some more comments to support the claim that "being in epistemic contact with one's conscious experiences is part of what it is to have them" (Kirk 2008, 87).[8]

So far so good. However, we need to explore the consequences of the supposed necessity of epistemic contact for Kirk's argument from its very start. Now I am going to argue that it blocks Kirk's argument *with respect to those cognitive mental states that stand in that necessary relation with experiences* already on its first stage (the argument for the incoherence of the e-qualia story).

Suppose that indeed it is conceptually impossible for there to be an experience and no cognitive state having that experience as its object. For simplicity sake, suppose that it is conceptually impossible for there to be an experience and no awareness of that experience. In that case, the relation between the experience and the awareness of this experience *is not that of causation* (causal links are contingent; they can always c-possibly be severed) but some special, *sui generis*, relation. Let us dub this relation as "superintegration". If so, there is no causation from the experience to its awareness but there is the awareness of the experience. Kirk's argument for the incoherence of the e-qualia story fails because it just does not take into account that there can be superintegration rather that causation between a non-physical experience and the awareness of that experience. In this case,

[8] This argument is mentioned already in (Kirk 2005, 50); however, there Kirk is more cautious about it and does not make it part of his argument: "although Chalmers's assumption is plausible, it is not needed for this argument" (Kirk 2005, 50). Instead, Kirk relies on the epistemic intimacy argument. In (Kirk 2008) things are reversed: Kirk omits the epistemic intimacy argument and relies on the argument from the necessity of epistemic contact with experiences.

an experience and the awareness of it are, in a sense, not two really distinct causally connected states, but two inseparable aspects of the same state (so that their separation is not even conceptually possible). Perhaps it is something like sides and angles of a polygon: although sides are not angles, it is even conceptually impossible for there to be the former without the latter, and it would make no sense to say that sides cause angles or *vice versa*.

Two things should be noted about this refutation of Kirk's argument for the incoherence of the e-qualia story.

First, it is available only for a zombist who admits that those cognitive mental states that are superintegrated with experiences are non-physical. That is, a zombist should be, in our terms, "Cartesian" *at least with respect to some cognitive mental states* (such as my present awareness of my present pain).

As for a *thoroughly* non-Cartesian zombist—that is, one who holds that all cognitive mental states (including such as my present awareness of my present pain) are physical—such a zombism is clearly incompatible with superintegration: if experiences are non-physical but my awareness of my experiences is physical, then the former and the latter are distinct and cannot be superintegrated. This can be used as an argument against the view that combines dualism with thorough cognitive physicalism. However, note that this argument is entirely independent from Kirk's anti-zombist argument; if it undermines the mentioned variety of dualism, it makes it on its own, and Kirk's anti-zombist argument does no job here. (Note that this outcome is just what Howard Robinson says in the remark quoted in subsection 3.1.)

On the other hand, as far as other varieties of dualism (wholly or partially "Cartesian"—those that admit that *at least some* of our cognitive mental states are non-physical) are concerned, the acceptance of the claim about superintegration invalidates Kirk's argument for the incoherence of the e-qualia story: if it is not causal link but superintegration that makes us aware of our experiences, and if this awareness is indeed not a distinct mental state but an aspect of experiences that cannot be c-possibly severed from them, then there is no contradiction between (E3*) that says that experiences (e-qualia) are causally inert and (E5) that says that there is an

epistemic contact with experiences—the epistemic contact is inbuilt in experiences themselves. [9]

The result is that far from saving Kirk's argument, the acceptance of the claim about superintegration blocks it at the first stage; at the same time, it undermines the view that combines dualism with thorough cognitive physicalism.

Second, the claim about superintegration is plausible only for the cases when an experience and a cognitive state directed at that experience are simultaneous. It may well be the case with my present-moment awareness of my present-moment experience. But it cannot be the case with my present-moment awareness (or thinking) of my a-day-ago or even a-few-moments-ago experience. If there is some temporal distance between an experience and a cognitive state having that experience as its object, there should necessarily be a causal link. If so, Kirk's argument can be run beyond its first stage (concerned with the coherence of the e-qualia story) only for those cognitive mental sates about experiences that are not superintegrated with the experiences they are about (such as my thinking about my past experiences). However, a zombist can successfully decline this argument as was explained in section 5.

The general outcome of this discussion is that there are two varieties of zombism that remain unscathed by Kirk's anti-zombist argument as well as by the claim about superintegration:

- a Cartesian dualism that holds that cognitive mental states are non-physical;
- a partially Cartesian interactionist dualism that holds that such states as my present awareness of my present experiences are non-

[9] At this point, the objection can be tried that the e-qualia story assumes cognitive physicalism, and of course, a Cartesian dualist should admit that *on that assumption*, the e-qualia story is incoherent. However, such an objection would be entirely beside the point. Of course, if we supplement the e-qualia story with the clause

(E6) All such states as being aware of experiences are physical,

then a dualist who accepts the claim about superintegration should agree that the e-qualia story+(E6) is incoherent. However, now to make his case, Kirk would be required to show that such a Cartesian dualist should admit the c-possibility of the world z*+(E6). I have no idea how he could do it.

physical, even if other cognitive mental states (such as my thinking about my past experiences or about non-experiential objects) are physical.

And the second part of Kirk's anti-zombist argument (having to do with the $z \rightarrow z^*$ transformation) achieves nothing at all.

References

Bailey, Andrew. 2009. "Zombies and Epiphenomenalism." *Dialogue* 48: 129–14. https://doi.org/10.1017/S0012217309090076

Chalmers, David. 1996. *The Conscious Mind.* New York: Oxford University Press.

Chalmers, David. 1997. "Moving Forward on the Problem of Consciousness." *Journal of Consciousness Studies* 4 (1): 3–46.

Chalmers, David. 1999. "Materialism and the Metaphysics of Modality." *Philosophy and Phenomenological research* 59 (2): 473–96. https://doi.org/10.2307/2653685

Chalmers, David. 2002. "Does Conceivability Entail Possibility?" In *Conceivability and Possibility*, edited by T. Gendler and J. Hawthorne, 145–200. New York: Oxford University Press.

Chalmers, David. 2004. "Imagination, indexicality, and intensions." *Philosophy and Phenomenological Research* 68: 182-90. https://doi.org/10.1111/j.1933-1592.2004.tb00334.x

Chalmers, David. 2010. "The Two-Dimensional Argument Against Materialism." In D. Chalmers, *The Character of Consciousness*, 141–205. New York: Oxford University Press.

Kirk, Robert. 1974a. "Sentience and Behaviour." *Mind* 83 (329): 43–60. https://doi.org/10.1093/mind/LXXXIII.329.43

Kirk, Robert. 1974b. "Zombies v. Materialists." *Proceedings of Aristotelian Society*, 48: 135–152.

Kirk, Robert. 2005. *Zombies and Consciousness.* Oxford, New York: Oxford University Press.

Kirk, Robert. 2008. "The Inconceivability of Yombies." *Philosophical Studies* 139: 73–89. https://doi.org/10.1007/s11098-007-9103-2

Kirk, Robert. 2017. *Robots, Zombies and Us. Understanding Consciousness.* London, New York: Bloomsbury Academic.

Levine, Joseph. 2001. *Purple Haze: The Puzzle of Consciousness.* New York: Oxford University Press.

Robinson, Howard. 2004. "Dennett on the Knowledge Argument." In *There's Something about Mary*, edited by P. Ludlow, Y. Nagasawa, and D. Stoljar, 69-73. Cambridge, London: The MIT Press.

Robinson, Howard. 2016. *From the Knowledge Argument to Mental Substance.* Cambridge University Press.

Organon F 29 (1) 2022: 109–127
https://doi.org/10.31577/orgf.2022.29105
ISSN 2585-7150 (online)
ISSN 1335-0668 (print)

RESEARCH ARTICLE

Factualism and Anti-Descriptivism: A Challenge to the Materialist Criterion of Fundamentality

Víctor Fernández Castro*

Received: 11 February 2020 / Revised: 27 January 2021/ Accepted: 3 June 2021

Abstract: Inspired by the work of Sellars, Cumpa (2014, 2018) and Buonomo (2021) have argued that we can evaluate our metaphysical proposals on fundamental categories in terms of their capacity for reconciling the scientific and the manifest image of the world. This criterion of fundamentality would allow us to settle the question of which categories among those proposed in the debate—e.g., substance, structure or facts—have a better explanatory value. The aim of this essay is to argue against a central assumption of the criterion: semantic descriptivism. Specifically, I aim at showing that the criterion rests on the idea that the manifest picture is mostly a description of the world, and thus, it commits us with certain realism. Instead, I argue that at least some of the vocabulary we use to construct our manifest picture of the world, mental vocabulary, is evaluative rather than descriptive and thus creates problems in reconcile the manifest picture with scientific psychology and neurosciences. I conclude with some remarks on alternatives that could provide a way out of the fundamentality criterion.

Keywords: Descriptivism; factualism; fundamental categories; mental vocabulary.

* Universidad de Granada
https://orcid.org/0000-0001-7627-5738
FiloLab-UGR, Universidad de Granada, 18011 Granada, Spain
vfernandezcastro@ugr.es

1. Introduction

In "Philosophy and the Scientific Image of Man" Wilfrid Sellars proposes, as the fundamental task of the philosophical endeavor, to reconcile the scientific image of the world with the manifest image, that is, the image produced by our scientific theories of natural sciences with the image we all take for granted. The difference between these two images turns to be especially evident when contrasting the common-sense perspective of humans as autonomous persons, responsible of their actions and motivated by thoughts and feelings, with a scientific perspective describing humans as biological systems without free will whose behavior is caused by physical processes and mechanisms.

Following in the footsteps of Sellars, Cumpa (2014, 2018) and Buonomo (2021) have proposed a scientific turn in metaphysics according to which we must assess our ontological theories regarding the fundamentality of the world on the basis of the contribution they can make to the reconciliation between the two images: *the materialist criterion of fundamentality*. As Buonomo (2021) puts it, "the scientific turn in metaphysics takes the fundamental categories to be the ones that play an essential role in the explanation of the relation between the ordinary world and the physical universe, providing us with a unified image of the world as a whole" (p. 795).

The aim of this article is to question an assumption of the materialist criterion of fundamentality which claims that the analysis of linguistic behavior necessarily produces, as a manifest image, a common-sense realism. This assumption, we argue, is based on a Cumpa's commitment to a form of semantic descriptivism (Chrisman 2007; Frapolli and Villanueva 2012; Gibbard 2003). Given that, we present a challenge to the criterion in conditional terms: If semantic descriptivism turns to be wrong about certain areas of discourse, and thus, the analysis of linguistic behavior cannot produce a manifest image that allow reconciliation, then the materialist criterion cannot be applied. In section 2, we present Cumpa's materialist criterion of fundamentality and how it works. In section 3, we present the central conditional argument and how it jeopardizes the criterion. Further, we argue that such argument puts defenders of the materialist criterion of fundamentality in a difficult situation. In section 4, we present several

arguments from non-descriptivism regarding mental vocabulary to emphasizes the impossibility of reconcile folk mentalism and psychology/neurosciences in metaphysical terms.

2. The materialist criterion of fundamentality

According to the traditional Aristotelian characterization of metaphysics, ontology is concerned with the study of being *qua being*, i.e., the identification of the most general and fundamental categories under which things fall and to characterize the relations between these categories. As such, ontology aims at seeking to understand the concept of being and existence and the properties and features that existence things exhibit as beings and existents. In this sense, a central function of metaphysics is to provide a map of the structure of all things. Now the question is what is the most fundamental of category of the world? Several contemporary authors (Heil 2013, Lowe 2011) have followed Aristotle in claiming that "substance" is the most fundamental category, while other authors have supported other categories like "structure" (French 2014, Ladyman and Ross 2007) or "facts" (Buonomo 2021, Cumpa 2014, 2018) as the fundamental category of the world.

The debate around the fundamentality of categories leads us to the question of how to decide between the different competing views (Cumpa 2020). Aristotle defended substantialism on the basis that the category of substance is prior, simpler and independent of other categories. For instance, substance is simpler than other categories because it cannot be divided in other categories and it is independent because it does not require other categories to exist. Those criteria are still object of debate in contemporary metaphysics (see Heil 2012, 2-5, 15; Armstrong 1997, 139-149). However, recently, Cumpa (2014) and Buonomo (2021) have proposed a different criterion, what they call the materialist criterion of world-fundamentality. According to this criterion, we must assess our ontological theories regarding the fundamentality of categories on the basis of the contribution they can make to the reconciliation between the scientific images of the world and the perceived or manifested image of the world. In other words, the explanatory power of our theories about categories must be evaluated in terms of

how the given categories can contribute to a better understanding of how the resulting image of how the world is according to our better scientific theories is compatible with the picture resulting from our ordinary experience.

Now, how should we understand such reconciliation? to see how, consider how Cumpa constructs his argument supporting the idea that factualism is better positioned for the reconciling task than other theories like substantialism. According to the criterion, a category must make sense of propositions of the type "A is F" where the two relata of the categorical scheme belong to different images. For a given proposition "the tomato is red", the categorial structure allows, for instance, the tomato to be located in the manifest image ("Tomato") and being red in the scientific image ("X is able to reflect a dominant wavelength measures between 618 and 780 nm"). According to Cumpa, factualism allows these type of propositions because being red is not a property of the tomato but a constituent of the fact, while substantialism does not make sense of such 'cross-sectional' propositions because the substance and its accident must be at the same level (Cumpa 2014, 321), that is, the categorial structure does not make the job as far as the structure 'S is P' corresponds either to the level of things (Gracia 1987) or to the elementary particles of physics (see Heil 2012, 52).

In brief, Buonomo and Cumpa argue that our disputes regarding the fundamentality of categories must arbitrated by an alternative criterion to those classically recognized in the debate like priority, simplicity or independency. According to this alternative criterion, a particular theory of categories is better than other when it can accommodate or make sense of propositions of the form "A is F" where the two relata can belong to different images. As a result, our better theories metaphysical theories will reconciliate our scientific understanding of the world with our everyday experience to the extent that our metaphysical categories can be compatible with propositions like "The tomato is able to reflect a dominant wavelength measures between 618 and 780 nm".

3. The challenge

As Heil (1998) said, "science does not speak with a single voice". There is no such thing as science, but sciences (physics, chemistry, meteorology,

geology, etc.) which focus on a strictly circumscribed domain. In this sense, it is inevitable that each science delimits the type of questions that are relevant, and that when certain limits are reached, the buck is passed to another science. However, even if each sciences were completely successful in accounting for its limits, it would still remain to evaluate how each science is pronounced in relation to each other, and of course, to our ordinary experience. In this sense, even if one doubts the value of pursuing a fundamental categorical scheme and embrace metaphysical pluralism, one can still find a certain value in the criterion of fundamentality as it would allow us to evaluate the relation of the metaphysical categories involved in a particular theory or science in relation to our everyday experience. Now, the fundamentality criterion requires that one of the stories belongs to the scientific sphere which is delimited by the particular science applied in the given domain. However, how can we define the area of application that belongs to the manifest image? What criterion do we use to decide what falls under the "ordinary level of thinghood with which ordinary people are acquainted in their commonsensical and practical experience" (Cumpa 2014, 319)?

Cumpa considers that the source of knowledge we must take into account in order to specify the ordinary world is not phenomenology, but rather the analysis of ordinary linguistic behavior. This analysis, he holds, leads us to what he calls 'common-sense realism'. Similarly, Buonomo claims that:

> "[c]ommonsense realism and scientific materialism represent the two methodological assumptions of the scientific turn in metaphysics. On the one hand, common sense realism accepts the ordinary level of thinghood we are acquainted with in our everyday lives and that we speak about in ordinary discourse. On the other hand, scientific materialism considers the scientific level of thinghood that scientists study through experimental research and represent with scientific theories." (Buonomo 2021, 796)

The analysis of linguistic behavior results into a manifest image that commits us with realism about the entities that populate our world. In this sense, the materialist criterion of fundamentality presupposes that the analysis of

our ordinary vocabulary will produce a common-sense image populated with objects, relations, and properties.

Notice that the scientific turn is undertaking an important semantic commitment to the idea that our everyday discourse is necessary descriptive. So, the materialist criterion, as specified by Cumpa and Buonomo, is based on semantic descriptivism. Descriptivism is the stance "whereby it's assumed that since semantic content of indicative sentences is standardly given in terms of their truth-conditions, the characteristic function of all indicative sentences is to describe worldly objects, properties, and relations" (Chrisman 2007, 227). In other words, the idea behind the characterization of the ordinary world relies on the assumption that the function of linguistic expressions is mainly descriptive. Certainly, this would seem obvious in the areas of discourse Cumpa and Buonomo are thinking of; for instance, ordinary objects (tables, chairs) and their properties (brown, rigid). However, this is not necessarily the case for all areas of discourse. In philosophical literature, we can find a set of views that share the denial of the descriptivist reading of a certain type of expressions or sentences. For instance, several views in metaethics like ethical expressivism (Gibbard 2003) or quasi-realism (Blackburn 1998) deny that sentences such as 'eating meat is wrong' describe a fact, namely, that a piece of behavior has a value property (being wrong). Similar positions are maintained about expressions such as epistemic attributions (Chrisman 2007, Field 2009), logical concepts (Brandom 2001), attribution of rationality (Gibbard 1990, Frapolli and Villanueva 2018) or modal expressions (Blackburn 1986, Thomasson 2014). The analysis of linguistic behavior, these authors suggest, can result in discovering that certain vocabulary is non-descriptive, and thus, its use does not commit us to any particular metaphysical counterpart.

The descriptive assumption of Cumpa and Buonomo have two important negative consequences for the criterion of fundamentality when seen from the perspective of the anti-descriptivist analysis. First, as several authors have argued (Chrisman 2008, Horgan and Timmons 1992, Mackie 1977), assuming descriptivism entails strong metaphysical commitments to the existence of nonnatural facts or entities like, for example, evil or goodness. These kinds of metaphysical entities, however, do not seem to be the kind of objects, properties and relationships that are part of a common-sense realism. In this sense,

the descriptive commitments of Cumpa and Buonomo do not seem compatible with the idea of the manifest image that they themselves promulgate. The descriptivist assumption produces an untenable manifest image that does not correspond with common-sense realism.

Second, if the anti-descriptivist analysis is right, the semantic analysis of natural languages that Cumpa endorses does not seems to produce even an image that is reconcilable with the scientific image when understood from the right metaphysical category. That is because if certain areas of discourse do not refer or state worldly aspects, then worldly metaphysical categories do not seem to apply to them. It is precisely this last consequence which seems especially challenging for Cumpa and Buonomo when we attempt to apply their criteria to certain areas of discourse that must be reconcilable with a scientific image but that are subject to an anti-descriptivist analysis like, for instance, mental vocabulary in relation with neurosciences or cognitive psychology.

To see how, consider that, according to Cumpa and Buonomo, the materialist criterion of fundamentality requires our categories to be able to explain how propositions involving the given terms can have their components in different images. The basic assumption is that propositions like "Pablo believes that the toy is on the table" must be understood from the a categorial structure that allows understand the two component of the proposition from common-sense realism: "a person named Pablo", "a particular mental state", and from the perspective of the scientific image like "A biological organism P", "a neuronal state M". Then, the Sellarsian question of ontology is to reconcile the tension that, for instance, "a mental state" and "a neural state M" does not seem to have the same properties but are really the same object (Cumpa 2018). The tension, Cumpa and Buonomo argue, is resolved when a particular category like "facts" allow to say that the propositions "A person named Pablo is in a particular mental state M" and "A biological organism P is in a neuronal state M" represent the same fact. Factualism can claim that both propositions represent the same fact precisely because we can exchange the relata to form two different propositions with one relata in each image but referring to the same fact: "A biological organism P is in a mental state M" and "A person named Pablo is in a particular neuronal state M". Now, the problem is that anti-descriptivist analysis of mental vocabulary does

not result into a common-sense realism. According to anti-descriptivism, the sentence "Pablo believes that the toy is on the table" does not describe or represent a particular object, property or relation. In particular, the expression "X Believes that the toy is on the table" cannot be substituted for an expression like "x is in a mental state M" where the expression represents or state for a property or a worldly aspect because the expression "believes" does not have descriptive meaning.

In a nutshell, metaphysical categories cannot help to reconciliate the two images because the linguistic analysis of the manifest image does not necessarily result into a picture where those categories apply. Now, such a claim just holds if anti-descriptivism of mental vocabulary turns to be right. But, what exactly mental vocabulary does if it does not describe? Do we have compelling arguments for supporting anti-descriptivism?

4. Anti-descriptivism and psychology

The point of contention raised in this paper is not straightforwardly tied to anti-descriptivism regarding modality or other metaphysical expressions (Blackburn 1986, Thomasson 2014). The key point is not whether expressions like "possibly, p" or "it is a fact that p" describe or not. On the contrary, the idea is that the criterion of fundamentality presupposes that the manifest image as produced by a linguistic analysis must be grounded in the reality in a way that every predicate or expression that compound a judgment of the manifest image is somehow anchored in the world; and thus, subject to be reconciliate with the scientific image[1]. However, we argue, if a descriptivism regarding mental vocabulary is right, and mental states predicates are not anchored in the world, the reconciliation is not possible, and thus, the criterion is useless, at least, for the domain of psychology and neurosciences in connection with the manifest image regarding our minds. In this section, we recapitulate some arguments supporting a non-descriptivist analysis of mental predicates.

[1] Thanks to an anonymous referee for pointing out the possibility that these two different projects could be confused.

Anti-descriptivism regarding mental states is a position that can be associated to different families of theories that goes from classical dispositionalism[2] of Wittgenstein (1953) and Ryle (1949) or the parentheticallism of Urmson (1952) to more contemporary theories like expressivism (Fernandez Castro 2017, Frapolli and Villanueva 2012, Perez-Navarro et al. 2019; Pinedo-García 2020), communicative conceptions of attribution (Fernandez Castro (2020), Tooming (2016), Van Cleave and Gauker 2010) or radical socio-cultural constructivism of mindreading (Almagro-Holgado and Fernandez Castro 2019; Fenici and Zawidzki 2020). Although they radically differ in the details, these views share the basic claim that mental states vocabulary serve for a different function than describing or tracking each other psychological states. For instance, Ryle (1949) understands dispositional terms[3] as inferential tickets: "an inference ticket (a season ticket)

[2] Although the work of Wittgenstein and Ryle is usually presented in contraposition to theories about the nature of the mind, like dualism or functionalism (see Ravencroft, 2005), Ryle and Wittgenstein present their views as positions about the use of psychological concepts, rather than views about the ontology of the mind. Moreover, Ryle and Wittgenstein do not have a realist interpretation of dispositional vocabulary, that is, they did not understand dispositional ascriptions as describing psychological states (Acero and Villanueva 2012, Freitag 2017, Glock 1996; Hacker 2010, Ter Hark 2001, Heras-Escribano and Pinedo-García 2018. Tanney 2007, 2009)

[3] Wittgenstein and Ryle systematically emphasize the idea that their research is not ontological but logical or conceptual. His philosophical enterprise is not to describe human psychological processes or to propose scientific theories concerning the mind: "The book does not profess to be a contribution to any science, not even to psychology. If any actual assertions are made in it, they are there through the author's confusion of mind" (1962, 196). On this account, the philosophical purpose of Ryle is to provide a conceptual clarification of how mental concepts are used, rather than elucidating what 'knowing', 'feeling' or 'remembering' is. Similar ideas can be found in Wittgenstein's work (1953, §89-90, 127, 199, 232, 392, 496, 574; 1974, 60). For instance, he claims: "Our investigation is therefore a grammatical one. Such an investigation sheds light on our problem by clearing misunderstandings away. Misunderstandings concerning the use of words, caused, among other things, by certain analogies between the forms of expression in different regions of language —Some of them can be removed by substituting one form of expression for another; this may be called an "analysis" of our forms of expression, for the process is sometimes like one of taking a thing apart" (Wittgenstein 1953, §90).

which licenses its possessors… to move from one assertion to another, to provide explanations of given facts, and to bring about desired states of affairs by manipulating what is found existing or happening" (p. 117). Expressions such as 'Sara believes that Riga is the capital of Latvia' function to make inferential moves: 'Sara believes that Latvia has a capital', 'If Sara wants to move to the capital of Latvia, she will take a flight to Riga' and so on. However, understanding dispositional terms as inferential tickets goes against considering them factual psychological states. As Tanney (2007, 2009; see also Heras-Escribano and Pinedo-García 2018) has emphasized, Ryle insists systematically in abandoning: "the preposterous assumption that every true or false statement either asserts or denies that a mentioned object or set of objects possesses a specified attribute" (Ryle 1949, 115).

Another example of how to understand non-descriptivism regarding psychological states is through their pragmatic function. Several authors argue that firs person ascriptions of mental states do not serve for describing one's mental states but for indicating certain degree of uncertainty or how to understand a particular proposition (Fenici & Zawidzki, 2020, Urmson 1952, Wierzbicka 2006). This means that, in sentences such as "I believe that the Indian restaurant is closed", the verb "believe" is not describing a mental state properly but merely indicating a low degree of commitment to the proposition "the Indian restaurant is closed". As Wierbicka (2006) points out, verbs in this use serve to modulate the interpretation of the proposition that falls under the scope of the verb. The verb "believe" serves to deny our knowledge of something, but not by saying "I don't know", but by saying "I don't say: I know". Similar analyses have been extended to third-person ascriptions (Fernandez Castro 2019, van Cleave and Gauker 2010, Geurt 2021), for instance, van Cleave and Gauker (2010) argue that third person ascriptions of desire, for instance, are used to carry out vicarious speech acts, so sentences like "Mom wants us to clean the room" serve to make a command (clean your room!) on the behalf of another person (the mother).

Be that as it may, the key point is that we have different analysis to motivate a non-descriptivist understanding of mental states predicates. Now, do we have arguments to support them? Ryle and Wittgenstein developed different argument to support non-descriptivism. For instance, Ryle

argues that mental dispositions, as skills, cannot be witnessed or captured, and thus, they are not metaphysically grounded in the world:

> Now a skill is not an act. It is therefore neither a witnessable nor an unwitnessable act. To recognise that a performance is an exercise of a skill is indeed to appreciate it in the light of a factor which could not be separately recorded by a camera. But the reason why the skill exercised in a performance cannot be separately recorded by a camera is not that it is an occult or ghostly happening, but that it is not a happening at all. (Ryle, 1949/2009, 22)

Skills, as other mental states, cannot be recorded with a camera, they are not winessable (or unwitnessable), not because they are hidden, but because they are not the type of mental phenomena we can point out or describe. In a similar vein, Wittgenstein (1967) presented the argument of duration, according to which, contrary to descriptive states, it does not make sense to say that a dispositional state (belief, desire, hope) takes time:

> Is "I hope ..." a description of a state of mind? A state of mind has duration. So "I have been hoping for the whole day" is such a description; but suppose I say to someone: "I hope you come"- what if he asks me "For how long have you been hoping that?" Is the answer "For as long as I've been saying so"? Supposing I had some answer or other to that question, would it not be quite irrelevant to the purpose of the words "I hope you'll come"? (Wittgenstein 1967, §78)

While it makes sense to ask for how long a state of affairs has been the case, it is unusual to ask for the duration of propositional attitude. Thus, the type of condition criteria of a propositional attitude ascription differs from those of a description. Another argument in that direction has to do with the grammatical or logical connection between a propositional attitude verb and its propositional object. When we say, 'Sara hopes that Beyoncé will record a new album', Wittgenstein argues, the established connection between the propositional object and the subject 'Sara' is not empirical, but logical (Wittgenstein 1967, 1974), and this type of connections cannot be described. The meaning of an expression is given by its connection with other expressions

(Wittgenstein 1974, §7). Thus, the connection between Sara's hope and its fulfillment is not empirical, and as such, is not descriptive. If Sara behaves in accordance with her hopes, we would say our attribution is right, and we would say is wrong otherwise; but this depends on the logical behavior of the concept 'hope' and not on an independent empirical connection between Sara and the proposition 'Beyoncé will record a new album'.

For the current purpose, another important argument lies on the impossibility of linking the vocabulary of sciences and the mental and appears on the work Donald Davidson (1970, 1991). Davidson presents different arguments supporting the claim that we cannot draw strict laws connecting the mental vocabulary and the vocabulary of physics. For instance, Davidson (1970, 172) suggests that we cannot establish strict laws between the mental discourse and the discourse of the physical sciences without changing the subject because the features of the two different vocabularies are unique to each one. As Ramberg (2000) has convincingly argued, this criterion does not apply uniquely to the distinction between the mental and the sciences, but also, to the distinction between physics and the special sciences. As he puts it: "Davidson grants that the relevant kind of law—that is, the strict kind – is no more likely to link special sciences to physics than it is to link psychology to physics" (p. 359). But, Davidson (1991) presents a distinctive reason for emphasizing the peculiarity of the mental vocabulary, i.e., the normative elements of mental states attributions. The critical question is not only that the vocabulary of agency involves the application of norms, but that the norms provide structure to the vocabulary (Ramberg 2000, 359). When we interpret others' actions, we are trying to find patterns by finding descriptions of what the other is doing. Finding such patterns depends on normative criteria of application of the concepts. In this sense, mental vocabulary may not differ from the vocabulary of sciences. However, Davidson's argue, finding such patterns requires taking a normative standpoint invoked by the charity principle, that is, we must assume that our interpretee meets the norms of rationality in order to find such patterns. Mental vocabulary does not only require norms of application but making claims about what sort of patterns count or not as mental. Thus, our interpretation of other creatures as mental are so intrinsically connected to a normative attitude that "If we were to drop the normative aspect from psychological explanations, they would no longer

serve the purposes they do. We have such a keen interest in the reasons for actions and other psychological phenomena that we are willing to settle for explanations that cannot be made to fit perfectly with the laws of physics" (Davidson 1991, 163). In Davidson's view, mental vocabulary serves a distinctive purpose than the vocabulary of sciences, a purpose that is not merely picking up objects for prediction and control. Mental vocabulary serves us to reveal the traits that allow us to recognize ourselves as creatures subject to moral and rational considerations, who can be burden with duties, commitments and rights (Ramberg 2000, 366).

Finally, several contemporary defenders of expressivism have defended that disagreement involving normative concepts also manifest an evaluative (and non-descriptive) function of those concepts (Chrisman 2007; Field 2009, Perez Navarro et al. 2019). According to those authors, disagreements involving normative concepts cannot be resolved by appealing to fact. In order to see the move, consider the following examples of disagreement:

[1] Shaq: The earth is flat
Kyrie: The earth is not flat

[2] Chris: Waterboarding is wrong
Hitch: Waterboarding is not wrong

Notice that the disagreement between Shaq and Kyrie can be solved by clearing up the relevant facts, viz. determining whether the earth is flat. Instead, the disagreement between Chris and Hitch does not necessarily dissolve after determining the relevant facts. We can imagine a situation where Chris and Hitch agree on all factual matters and still disagree about whether waterboarding is wrong. Moreover, the disagreement in question does not necessarily dissolve when the normative standards are made explicit, removing the possibility that description is dependent of norms:

[2]' Chris: According to the Human Rights Declaration, waterboarding is wrong

Hitch: According to the Eight Amendment, waterboarding is not wrong

We can conceive situations where Chris and Hitch do not necessarily resolve their dispute after making the norms explicit. Now, Perez-Navarro et al.

(2019) have elaborated upon this argument to defend that we can identify evaluative disagreements involving belief attributions. They illustrate the point with an example by Dennett (1978) where he invites to consider the case of Sam, an art critic who has promoted the paintings of his son. There are two possible interpretations of this situation: "a) Sam does not believe the paintings are any good, but out of loyalty and love he does this to help his son, or (b) Sam's love for his son has blinded him to the faults of the paintings, and he actually believes they are good" (Dennett 1978, 39). Now, suppose for the sake of the argument that there exists a reliable way of determining the cause of someone's actions. Imagine, as Dennett does, that we have the technology to write a specific judgment in Sam's brain. Imagine that we write 'my son's paintings are great' at the moment he is promoting his son's paintings. In fact, we can suppose that this was the occurrent cause of the action (promoting his son) at that moment. Dennett's point is that, even in this extreme case, there are no deep facts we can appeal to in order to decide whether the ascription of this belief is certainly explanatory of the situation. Someone could examine the past and future circumstances of Sam and suspend the interpretation that Sam believes that his son's paintings are good. The interpreter could examine Sam's past behavior and realize that he systematically avoided assessing his son's paintings using the same aesthetics standards that he used for other artists, or that his subsequent behavior is incoherent with the decision of promoting his son's paintings. These circumstances would provide the interpreter with reasons to change his verdict. At the same time, the other interpreter could insist that the accurate ascription is the one that identifies the real cause of the behavior. However, it is dubious whether we can decide which belief ascription is right by appealing to the mere facts. Both interpreters could agree on all the relevant facts and differ on their ascriptions. Moreover, even when if both interpreters would make their norms of interpretation explicit—e.g., appealing to the Sam's incoherence or sincerity—the disagreement would not necessarily disappear.

As a result, we have reasons to believe that mental states attributions and predicates do not describe entities of any type. Acknowledging the possibility that linguistic expressions might not identify a particular object, relation or property may jeopardize the idea of a manifest image in common sense realist terms or a manifest image at all, and thus, the applicability of the materialist

criterion of fundamentality may be severely restricted. Certainly, this conclusion is dependent on the persuasion of non-descriptivist arguments. However, to the extent that the materialist criterion depends on a descriptive semantics, one should, at least, critically face the arguments and motivations behind non-descriptive semantics to save the applicability of the criterion.

5. Concluding Remarks

Where does this leave the scientific turn in metaphysics? one possible way to save the criteria is by finding an alternative possibility to ground the ordinary level of thinghood other than linguistic analysis. After all, it seems plausible to maintain that, even if the use of certain expressions is not aimed to describe or represent the world, ordinary people could have some common-realist intuitions concerning the status of our mental life. Now, the question is whether we could find a way to rescue these intuitions. Certainly, one possible alternative is to appeal to phenomenology as a way of constructing the manifest image but Cumpa seems inclined to resist such a strategy (Cumpa 2014, 320). Although he does not specify why, one may speculate that the reason is related to the possible problems one may encounter when trusting one's own experience or intuitions regarding mental states; for instance, the possibility that our own experience dramatically differs from each other's. A plausible middle path could try to exploit Dennett's (1991) *hetero-phenomenology*. In this view, we could create a profile of the people's reports about their own experiences and intuitions regarding other's and their own mental states (Dennett 1991, 76-77). So, the ordinary level of thinghood could be grounded in people's reports about their own experience. We can control the problems emerging with phenomenology by testing only the intuitions that are statistically significant inside of a given population. Be that as it may, this alternative implies abandoning the analysis of linguistic behavior as the procedure to construct our manifest image.

Leaving aside the alternative in (hetero)phenomenological terms, there seems to be a deeper problem with Cumpa and Buonomo's reconciliatory project. In principle, as the history of science has demonstrated, it seems

likely that some important aspects of our manifest image do not lend themselves to a metaphysical reconciliation of some kind with science, but to another kind of assimilation such as the elimination or, like the case of the mind, a more complex assimilation than mere metaphysical mapping. Perhaps, the response can be found in the work of Sellars himself. Sellars (1956) seems to defend certain type of non-descriptivism when he says 'in characterizing an episode or a state as that of knowing, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says' (§36). Such a claim, along with his complaint that such descriptive treatment of knowledge would be 'a mistake of a piece with so-called 'naturalistic fallacy' in ethics" (§5), must be regarded as an indicator of the limits of the reconciling metaphysical enterprise; at least, if we understand the enterprise as establishing metaphysical connections between two types of images that serve radically different objectives and interests. The alternative may not be necessarily the skepticism, but simply seeking reconciliation beyond the metaphysical enterprise of categories, for example, understanding that the scientific picture must give us an adequate picture of how we humans, as natural beings, are able to create for ourselves a picture of the world that is presented to us in such and such a way. To try to assimilate one image to the other in terms of worldly categories is perhaps only a metaphysical dream.

Acknowledgements

The author would like to thank Manuel de Pinedo, Javier Cumpa and all the participants of the First Meeting of Physis (UCM, Madrid) and Filosofía y Análisis (UGR, Granada) in Analytic Philosophy at the Complutense University at Madrid for their valuable comments and suggestions. The study was supported by the projects PID2019-108870GB-I00 and PID2019-109764RB-100 of the Spanish Ministry of Science.

References

Acero Fernández, Juan José & Neftalí Villanueva Fernandez. 2012. "Wittgenstein's anti-descriptivism." In *VII SOLOFICI, Conference Proceedings*, edited by C.

Martínez Vidal, J.L. Falguera, J. M. Sagüillo, V. Verdejo, & M. Pereira-Fariña, 102–9. Spain: Servicio de Publicaciones de la USC.

Almagro Holgado, Manuel & Castro, Víctor Fernández. 2020. "The Social Cover View: a Non-epistemic Approach to Mindreading". *Philosophia* 48(2): 483–505. https://doi.org/10.1007/s11406-019-00096-2

Armstrong, David M. 1997. *A World of States of Affairs*. Cambridge: Cambridge University Press.

Blackburn, Simon. 1986. "Morals and Modals." In *Fact, Science and Morality,* edited by G. MacDonald, 119–41. Oxford: Blackwell.

Blackburn, Simon. 1998. *Ruling Passions*. Oxford: Oxford University Press.

Buonomo, Valerio. 2021. "The Scientific Turn in Metaphysics: A Factualist Approach." *Synthese* 198: 793–807. https://doi.org/10.1007/s11229-017-1445-5

Chrisman, Mathew. 2007. "From Epistemic Contextualism to Epistemic Expressivism." *Philosophical Studies* 135 (2): 225–54. https://doi.org/10.1007/s11098-005-2012-3

Chrisman, Mathew. 2008. "Expressivism, Inferentialism, and Saving the Debate." *Philosophy and Phenomenological Research* 77(2): 334–58. https://doi.org/10.1111/j.1933-1592.2008.00194.x

Cumpa, Javier. 2014. "A Materialist Criterion of Fundamentality." *American Philosophical Quarterly* 51(4): 319–24.

Cumpa, Javier. 2018. "Factualism and the Scientific Image." *International Journal of Philosophical Studies* 26(5): 669–78. https://doi.org/10.1080/09672559.2018.1533713

Cumpa, Javier. 2020. "Categories." *Philosophy Compass* 15 (1): 126–46. https://doi.org/10.1111/phc3.12646

Davidson, Donald 1980/2001. "Mental Events." In *Essays on Actions and Events*, 170–86. Oxford: Clarendon Press.

Davidson, Donald. 1991. "Three Varieties of Knowledge." *Royal Institute of Philosophy Supplements* 30: 153–66. https://doi.org/10.1017/S1358246100007748

Dennett, Daniel C. 1991. *Consciousness Explained*. New York: Little, Brown.

Dennett, Daniel C. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge: Bradford Books.

Fernandez Castro, Victor. 2019. "Justification, Conversation and Folk Psychology." *Theoria*. 34 (1): 73–88. https://doi.org/10.1387/theoria.18022

Fernandez Castro, Victor. 2017. "The Expressive Function of Folk Psychology." *Unisinos*, 18 (1): 36–46. https://doi.org/10.4013/fsu.2017.181.05

Field, Hatry. 2009. "Epistemology Without Metaphysics." *Philosophical Studies* 143 (2): 249–90. https://doi.org/10.1007/s11098-009-9338-1

Finkelstein, David (2003). Expression and the Inner. Cambridge: Harvard University Press.

Fenici, Marco & Tadeusz W. Zawidzki. 2020. "The Origins of Mindreading: How Interpretive Socio-cognitive Practices Get Off the Ground" *Synthese*. https://doi.org/10.1007/s11229-020-02577-4

Frapolli, María José & Neftalí Villanueva Fernández. 2012. "Minimal Expressivism." *Dialectica* 66 (4): 471–87. https://doi.org/10.1111/1746-8361.12000

Frapolli, María José & Neftalí, Villanueva Fernández. 2018. "Minimal Expressivism and the Meaning of Practical Rationality." In *Rationality and Decision Making,* edited by M. Hetmański, 1–22. Poznan, Poland: Brill Rodopi.

Freitag, Wolfgang. 2017. "Wittgenstein on "I believe"." *Grazer Philosophische Studien*, 95(1): 54–69. https://doi.org/10.1163/18756735-000018

French, Steven. 2014. *The Structure of the World: Metaphysics and Representation*. Oxford: Oxford University Press.

Geurts, Bart (2021). "First Saying, Then Believing: the Pragmatic Roots of Folk Psychology". *Mind & Language* 36 (4): 515–32. https://doi.org/10.1111/mila.12345

Gibbard, Alan. 1990. *Wise Choices, Apt Feelings: A theory of Normative Judgment*. Cambridge: Harvard University Press.

Gibbard, Alan. 2003. *Thinking How to Live. Cambridge*. MA, USA: Harvard University Press.

Glock, Hans. 1996. *A Wittgenstein Dictionary*. Oxford: Blackwell.

Gracia, Jorge. 1987. *Individuality: An Essay on the Foundations of Metaphysics*. Albany State University of New York Press.

Hacker, Peter M.S. 2010. *Wittgenstein: Comparisons and Context*. Oxford: Oxford University Press.

Heil, John. 2012. *The Universe as We Find It*. Oxford: Oxford University Press.

Heras-Escribano, Manuel & Manuel Pinedo-García. 2018. "Naturalism, Non-factualism, and Normative Situated Behavior." *South African Journal of Philosophy* 37 (1): 80–98. https://doi.org/10.1080/02580136.2017.1422633

Horgan, Terence & Mark Timmons. 1992. "Troubles For New Wave Moral Semantics: the Open Question Argument' Revived." *Philosophical Papers* 21(3): 153–75. https://doi.org/10.1080/05568649209506380

Ladyman, James, & Don Ross. 2007. *Everything Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Lowe, Edward J. 2011. "A Neo-Aristotelian Substance Ontology: Neither Relational nor Constituent." In *Contemporary Aristotelian Metaphysics*, edited by E. Tahko, 229–48. Cambridge: Cambridge University Press.

Mackie, John L. 1977. *Ethics: Inventing Right and Wrong*. London: Penguin Books.

Pérez Navarro, Eduardo, Castro, Víctor Fernández, González de Prado, Javier & Heras-Escribano, Manuel. 2019. "Not Expressivist Enough: Normative Disagreement about Belief Attribution." *Res Philosophica* 96 (4): 409–30. https://doi.org/10.11612/resphil.1794

Pinedo-García, Manuel. 2018. "Ecological Psychology and Enactivism: A Normative Way Out from Ontological Dilemmas." *Frontiers in Psychology* 11. https://doi.org/10.3389/fpsyg.2020.01637

Ramberg, Bjørn. 2000. "Post-ontological Philosophy of Mind: Rorty Versus Davidson." In *Rorty and his Critics* edited by R. Brandon, 351–69. Mass: Blackwell Publishers.

Ravencroft, Ian. 2005. *Philosophy of Mind: A Beginner's Guide*. Oxford: Oxford University press.

Sellars, Wilfried. 1963. "Philosophy and the Scientific Image of Man." In *Frontiers of Science and Philosophy* edited by R. Colodny, 35–78. Pittsburgh: University of Pittsburgh Press.

Tanney, Julia. 2009. "Rethinking Ryle: A Critical Discussion of The Concept of Mind." In *The Concept of the Mind 60th Anniversary Edition,* ix-lix. London: Routledge.

Tanney, Julia 2013. *Rules, Reason and Self-Knowledge*. London: Harvard University Press.

Ter Hark, Michel R. M. 2001. "Wittgenstein and Dennett on Patterns." In *Wittgenstein and Contemporary Philosophy of Mind* edited by S. Schroeder, 85–104. Basingstoke: Palgrave MacMillan.

Thomasson, Amie L. 2014. *Ontology Made Esy*. Oxford: Oxford University Press

Tooming, Uku. 2016. "Beliefs and Desires: From Attribution to Evaluation" *Philosophia* 45(1): 359–39. https://doi.org/10.1007/s11406-016-9756-1

Urmson, James O. 1952. "Parenthetical Verbs." *Mind*, 6(244): 480–96. https://doi.org/10.1093/mind/LXI.244.480

Van Cleave, Matthew & Christopher Gauker. 2010. "Linguistic Practice and False-Belief Tasks." *Mind and Language* 25 (3): 298–328. 10.1111/J.1468-0017.2010.01391.X

Wittgenstein, Ludwig. 1953/2001. *Philosophical Investigations*. Revised English translation by G.E.M. Anscombe, Oxford: Blackwell.

Wittgenstein, Ludwig. 1967. *Zettel*, Oxford: Blackwell.

Wittgenstein, Ludwig. 1974. *Philosophical Grammar*.Oxford: Blackwell

Wierzbicka, Anna. 2006. *English: Meaning and Culture*. New York: Oxford University Press.

Organon F 29 (1) 2022: 128–164
https://doi.org/10.31577/orgf.2022.29106

ISSN 2585-7150 (online)
ISSN 1335-0668 (print)

RESEARCH ARTICLE

Is There an Alternative to Moderate Scientism?

Szymon Makuła*

Received: 13 January 2020 / Revised: 11 January 2021 / Accepted: 20 June 2021

Abstract: This paper's primary purpose is to show that there is a peculiar alternative to scientism whose central thesis is not about sources of knowledge or the existence of various objects, but it aims at setting out a strategy to help decide which of the two mutually exclusive beliefs is the better one to adopt. Scientophilia, to coin a term, recommends preferring, without any discussion, a position consistent with the consensus of credible and reliable experts in a given domain. In case there is no such agreement, mainly because peers disagree with each other, or experts are difficult to identify, it is recommended for a scientophile to suspend judgment. Scientophilia is not a position on science or human knowledge boundaries, but it deals with the practical side of belief change. Verdicts made by this approach are partially similar to those offered by mild scientism, as scientophilia puts scientific knowledge as one of the most reliable sources. However, it is also consistent with mild antiscientism, as in some particular cases (for example, Moorean truths), it assigns reliable expertise to non-scientific experts. Therefore it is a third way.

Keywords: Antiscientism; demarcation of science; scientism; scientophilia.

* University of Silesia
https://orcid.org/0000-0001-6421-2748
Bankowa 11 40-007 Katowice University of Silesia, Poland
szymon.makula@us.edu.pl

1. What science is?

Before anything is said about scientism and scientophilia, it is first necessary to discuss the very concept of science. The question "what is science?" is similar to the Augustinian question about what time is. Until we make an effort to find an answer, the issue seems to be simple. However, once we try to take on this seemingly trivial challenge, we notice that we are dealing with an extraordinarily complicated and multidimensional human activity, one which is continuously evolving and changing. The term "science" has very positive connotations and suggests that we are dealing with the highest quality of knowledge. This word is so well-established in our language that, as Susan Haack notes, it often has an ennobling function (Haack, 2012). The prestige that goes hand in hand with this term is undoubtedly related to the natural sciences' success. The aura of reliability surrounding the word "science" gives rise to a strong temptation to use it for persuasion. We observe such attempts every day. Advertisements cite scientific research, the results of which assure us of the positive characteristics of the product offered to us. Participants in television debates willingly use the authority of science to authenticate their position in a dispute. Even university circles are not immune: various disciplines and fields of study containing the word "science" in their names, such as cognitive science, political science, social science, are proliferating at universities around the world (Haack, 2012). The variety of contexts in which the term "science" occurs raises the question of its exact definition.

Unfortunately, the very concept of science is vague (Hansson, 2013); therefore, it cannot be precisely defined. Nonetheless, the term "science" can be applied in the vast majority of cases, albeit the existence of a grey area, in which the use of this term will be ambiguous, is inevitable. No one will argue with the claim that tying shoes, playing tennis, or watching a film at the cinema are not examples of doing science. Similarly, no one will deny that in the Michelson–Morley experiment or the Hershey–Chase experiment, we are dealing with science *par excellence*. It does not mean that a scientific approach to tying shoes, playing tennis, or watching a film is impossible; one could not be more wrong, though in most cases, tying shoes is nothing other than tying shoes.

On the other hand, one can discuss whether a Michelson–Morley experiment finalized during a theatrical performance is scientific or not. Such discussions only begin when a particular issue is located in the grey area. As if that was not enough, not only do we not know where science begins and where it ends, but we do not know precisely when it was created. Was Aristotle's inquiry about the natural world a science or not yet? Did Ptolemy conduct science? Is Ibn al-Haytham's *Optics* a scientific work? Following Massimo Pigliucci (2017), it could be argued that they are in a sense, but certainly not in the way we talk about biology, astronomy, or optics today. As Robin Dunbar (1996) notes, it can be said that the traditional Japanese method of *ayu* fishing also has something scientific about it, as its effectiveness was because fishers managed to correctly recognize the mating habits of this species. This example does not mean that centuries ago in Japan, fishers practiced ethology in the same way it is practiced at modern universities.

Contrary to common opinion, there is no single scientific method; therefore, science's diversity is also manifested in its methodologies (Haack, 2016; van Woudenberg, 2011). Not all sciences are experimental, and not all sciences predict phenomena and explain them. The same holds true for using statistical methods, creating computer models, or using surveys. The reasons for this state of affairs can vary. Some scientific disciplines do not need specific tools, e.g., physicists will never use a survey in their work. Some fields cannot use specific research methods for various reasons, or they are applicable in only a minimal range; for example, in disciplines such as psychology or medicine, it is not always possible to use experimental methods due to ethical issues.

For these reasons, it cannot be said that there is such a thing as one science. There is a whole cosmos of sciences similar in some respects and different in others (Haack, 2016). Whenever we talk about science, we assume a definition that involves arbitrary decisions and includes disciplines that someone else would not include among the sciences. This does not mean that one can use the term "science" freely and, for example, put magic on a par with astronomy, as Feyerabend (1993) did in his "Against method". Similarly, as there is no one science, there cannot be one nonscience. It is worth noting that the term "nonscience" applies to a relatively

broad concept, which includes not only tying shoes, swimming, dancing, or watching movies, but also religion, practical knowledge, political beliefs, poetry, and a vast range of so-called pseudoscientific theories such as astrology, creationism, homeopathy, Lysenkoism, or phrenology. Every pseudoscientific theory should be classified as a nonscience, but not the other way around. A pseudoscientific claim is not only nonscientific but, contrary to other nonscientific activities, its proponent aims to create the impression that we are facing the most reliable knowledge in this particular subject matter, which is not the case. It follows that drawing a demarcation line between science and nonscience is more complicated than it might seem because the world of nonscience is internally diverse. That is why there are still fierce arguments about where the demarcation line separating science from nonscience should be placed (Hansson, 2013; Nickles, 2013; Pigliucci, 2013; Simonton, 2018).

2. Nonscience and pseudoscience

The problem of demarcating science from pseudoscience attracted philosophers of science's attention in particular and is often taken as equivalent to the more general term "demarcation of science." However, this issue is vital; not every demarcation line will be established to distinguish between science and pseudoscience, as it is essential to differentiate among other above-mentioned nonscientific activities. Indeed, one issue related to the distinction between science and pseudoscience can be generalized to the whole question of demarcation. As Hansson (2013) put it, "For a scientist distinguishing between science and pseudoscience is much like riding a bicycle" even though there is no explicit criterium of demarcation, that is to say, it is instead a matter of tacit knowledge. In most cases, most scientists will unanimously recognize scientific inquiries, just as most people will recognize a short man.

Moreover, just as it will be hard for us to pinpoint when a person ceases to be short, it will be hard for scientists to pinpoint the moment when a human activity begins or ceases to be science. This quandary also applies to both nonscience and pseudoscience. It is not the job of a layman to distinguish science from nonscience, but it is a proper task for an expert in

the field. There is no better candidate even if such an expert cannot establish a sharp boundary between them.

Nonetheless, it is crucial to distinguish between the broad and narrow meaning of the "science" term. The latter originated in the nineteenth century, and its meaning was restrained to the very study of nature. Science in a broad sense is a quest for seeking knowledge about the laws of nature and an attempt to discover, explain, and understand the mechanisms regulating and influencing our organisms' functioning, psyche, or the regularities governing social life. There is no reason why we should exclude psychology, sociology, economy, or history from the set of science in the aforementioned broad sense. These are also remarkable, methodically conducted, and academically acclaimed inquiries, and what is even more relevant, social sciences strive to produce the most epistemically warranted knowledge there is. If we consider epistemological success and reliability, then there is a meaningful discrepancy between natural sciences and social science, and the advantage of the former over the latter is indisputable. The credibility of the evidence and the verifiability of theses provided by climate science is incomparably more significant than those present in historical sciences. However, since historians have established the exact course of the events of the Holocaust, there is no reason to dismiss their claims merely because history is a less credible science than climate physics, especially since there are no other reliable studies on past events than those conducted by academic historians.

To sum it up, it is impossible to indicate the boundaries of the term "science", which means that it is impossible to indicate exactly where it begins or where it ends; it is also problematic to distinguish its elements (Blackford, 2017), and it will also be dubious about differentiating between science and nonscience and science and pseudoscience conspicuously. This does not mean that everything can be contained in this term, but that the powerful feature, namely scientificity, is a gradable property and can sometimes be overlooked or mistakenly attributed to an object. In the broad sense, adopted here, science is a conglomerate of many disciplines that intersect and mix. Indeed, science is not the only source of knowledge, but it is a recognized source (de Ridder, 2018). This recognition is based on the power of authority that we assign to science, which in practice is equivalent

to accepting scientific assertions. Not every scientific discipline stands out to the same extent as physics or the natural sciences in general, so not every science should be treated as an authority with the same clout as the natural sciences, and if any specific issue turns out to be a nonscientific one, it does not mean it is worthless. In some disciplines, especially in the social sciences, experts' opinions will be only slightly better than that of laypeople.

3. Science from a social epistemology perspective

If we look at science from an epistemological perspective, it is hard to deny that scientific theorems deserve proper respect because of the sciences considerable cognitive success. It does not mean that they should be treated as absolute truths and scientists as their infallible preachers (Haack, 2007). Nothing could be more wrong: scientific knowledge is fallible, uncertain, and far from perfect, like any other human creation. It is not the degree of certainty of scientific statements that deserves esteem, but the way scientists have succeeded in developing our ordinary ways of thinking.

Scientists engage in such activities as experiments, take measurements, collect data, analyze them, draw conclusions from them, publish in peer-review journals, compare their results, replicate their colleagues' studies looking for errors in them (Goldman, 1999). To this end, researchers are developing various standardized procedures that facilitate their evaluation. Unfortunately, it has not been possible to work out one universal recipe that would allow us to assess, always and everywhere, what evidence is needed to resolve a given dispute. Humanity is continuously improving old methods or developing new methods, and this work better in a given context but do worse in others. Some specific standards are common to many disciplines; others can be found only in mathematics, physics, yet others in psychology. Some disciplines have stringent and precise rules of evidence—here, the model is mainly formal sciences. In others, a lot still depends on the researchers and their decisions, as in the social sciences. Some pieces of evidence are so complicated that professionals need years to detect errors in them. Even in the case of mathematics, it can take years to detect an error in a proof. It took 11 years to find a fault in one of the famous four-color theorem's alleged proofs. It took decades to establish the consensus

which is currently adopted on the observed climate change. As early as 1991, 67% of the scientifically active climatologists were convinced that it is a human activity that is causing the planet to overheat; it was only in 2009 that this percentage approached 100% (it was about 97%) (Cook et al., 2016). Unfortunately, expert opinions are not always unanimous; there is much discussion in science, controversial views, unsolved problems. That is because the research and data collected do not always allow for an unequivocal adoption of a given conclusion. As if that was not enough, it is not uncommon for scientists to make mistakes or even to commit ordinary forgery (Fanelli, 2009). Additionally, there are phenomena such as merchants of doubt i.e., scientists paid by various interest groups who question the research results unfavorable to the client (e.g., the harmfulness of tobacco or the ecological effects of burning fossil fuels) (Oreskes and Conway, 2010).

Science defends itself against such problems quite a simple way: the scientific community is continuously keeping an eye on itself. It is possible thanks to various protective mechanisms, such as the blind review method of publications, replication studies, discussions, meta-analyses, new measurement methods, and tools. The foundation for these safeguards is the constant pursuit of processes, let us call them knowledge-making processes, which are as transparent and replicable as possible. Adaptation of these procedures means that in an ideal situation, every competent researcher should replicate, step by step, an experiment conducted by a colleague or replicate a measurement, thus checking whether the same result can be obtained. Various sciences manage to implement this idea to varying degrees, which does not change the fact that the pursuit of intersubjective communication is the common denominator of all kinds of scientific thinking. The safeguarding system, which results from implementing the idea of intersubjectivity to varying degrees, is far from perfect and is unable to stop us from finding errors and mistakes in science. In this case, the only thing we can hope for is to reduce their number.

4. Experts and nonexperts

The word "expert" has a broad meaning, as we commonly refer to people who have acquired an exceptional level of some skill or ability. In this sense,

an expert is both a chess player, car mechanic, ballerina, and volleyball player. However, for epistemological purposes, one's expertise is narrowed down to a cognitive extent; therefore expert is a person who not only possesses a substantial body of truths in a given domain but also she is sufficiently competent to form the right answers to new questions in her domain (Goldman, 2011).

Everyone would be a cognitive expert in an ideal world, and everyone could assess the quality of evidence behind two contradictory claims. Unfortunately, we do not live in such a reality; we have to deal with the fact that everyone is a layperson in our world. Even if someone is a nuclear physics professor, they are most likely a layperson in any other discipline such as crowd sociology, cognitive psychology, evolution theory, horse riding, chess gambits, fuzzy logic, mating habits of orangutans or ancient Roman law. It does not mean that we are all ignorant, but that even the greatest erudite will have achieved mastery of a few disciplines at most. Nobody will ever know all spheres of science, literature, music, sport, history, or philosophy. In other words, there is a significant division of labor in science (D'Agostino, 2016).

Specialization requires time and sacrifice, mostly when we talk about the natural sciences. They are characterized by such a high degree of complexity that a layperson would not understand even an abstract of a paper without proper training. The development of scientific disciplines and the following specialization have gone so far that the evaluation of evidence collected by experts is beyond a dilettante's capabilities or even for a single expert. In the last century, Derek de Solla Price observed a rapidly growing multi-author publication trend in science (Price, 1963). Nowadays, this tendency is even more visible, as the "Multi-authorship and research analytics" report claims, the most frequent number of authors is three, and the count of papers with at least 100 involved scientists is continuously growing (Adams et al., 2019). These scientists often represent different disciplines, which means that a single expert cannot even review interdisciplinary teams publications because his expertise is too narrow.

Besides our lack of competence, there is another reason we are doomed to scientists, which is that we have too little time. The continual increase in knowledge, measured by the number of scientific publications, is too vast

for one person to be able to take all of it in. In 2012, the number of annual publications exceeded 1.8 million (Ware and Mabe, 2012). It may be comforting to know that there are also such issues whose complexity level is so low that we cannot say there are laypeople in their case. Each of us is an expert in matters such as our pocket contents, our parents' names, or our place of residence; we know perfectly well whether we have a toothache or not. When someone asks us if we have a lighter or wants to know what time it is, we will not consult an expert because we can answer these questions immediately or know how to answer ourselves. In other words, in the face of such issues, we can trust ourselves. Unfortunately, we will not always have such comfort.

A person who thinks that she can decide whether human activity is the cause of the climate catastrophe, whether vaccines cause autism, whether GMO is harmful, or whether 5G technology harms the human brain is under an illusion. Our autonomy in such complicated matters boils down to merely recognizing some sources of information we have found credible and rejecting others as unreliable. There is always a third way, i.e., to suspend judgment. Unfortunately, such a skeptical attitude cannot save us from all dilemmas, because in some cases, the suspension of judgment is tantamount to taking action consistent with one of the disputed positions. A layperson may recognize that the dispute over climate change's genesis is overwhelming and thus refrain from taking a position. Unfortunately, the dispute over climate change also applies to human actions because one side recommends reducing CO2 emissions, and the other claims that such actions are unnecessary. Depending on whether a layperson will try to reduce their impact on the environment or not, they will act as recommended by one or the other party to the dispute. At least in some cases, we will not escape having to decide on whom to believe. That is why it is worth finding a strategy that gives one the least chance of making a mistake. The choice of such tactics is essential in the modern world of information overload. On the internet, one can find everything from scientific research, through reports about mermaids living in the Atlantic Ocean, to video recordings of alleged time travelers. The conclusion is as follows: we are laymen; hence the dependence on an expert's testimony is inevitable (Goldberg, 2016; Lackey, 2011); if so, our ability to assess the degree of expertise of others and their credibility is a crucial skill.

5. Informed trust in expert opinion

Regardless of whether we are talking about the natural or social sciences, scientific knowledge is the product of a complex structure built by large teams of people, and a layman's confidence in scientific claims can only be based on trust in these structures. However, contrary to John Hardwig (1991), trust in science does not have to be blind; more so, it could be, as Naomi Oreskes (2019) calls it, informed trust. Jennifer Lackey (2011), one of the so-called social epistemologist, notes that everything we know is more or less based on other people's testimony; undeniably, we are told such things as how everything around us works, what is going on in foreign countries, where our food came from, what is it made of, what happened before our birth. It is hard to pinpoint any specific part of our knowledge that we established without trusting in someone else testimony. It is precisely the same case with scientific discoveries. Problems start to emerge when we face many contradictory statements that express these testimonies, which is, unfortunately, an inevitable situation with scientific knowledge.

Social epistemology is an expanding philosophical discipline that offers some guidance in this baffling situation. Unlike plain epistemology, this very discipline is concerned not with abstract and theoretical issues but mainly with such practical problems as testimony, judgment aggregation, and peer disagreement. It is worth pinpointing that the following heuristic's primary purpose is to make our decision process about trusting in expert testimony more efficient. It is not designed to advise on such issues as establishing scientific truth, and its character is probabilistic, which means if a layperson follows these cues, she will increase the odds that she chose a reliable opinion.

Scientific experts differ from laypeople in several significant respects, including their extensive and substantive knowledge in a given field, and that they gravitate towards using this knowledge to answer new questions and solve current problems in their field, evaluate evidence gathered by their peers (Goldman, 2011). It is reasonable to treat an expert's opinion in their field of expertise as more reliable than that of a layperson because expertise in a particular field carries with it a specific type of authority, namely, cognitive authority. Of course, relying on such authority is fallible

under the fallible nature of scientific inquiry itself, but the layperson has nothing better up their sleeve, as it was concluded earlier.

There is no distinct point beyond which the layperson becomes an expert. Expertise is a continuous trait in which sheer ignorance lies on one side of the spectrum and extraordinary competence on the opposite. There is a consistent pattern, layperson or even a novice exploring a given domain of scientific knowledge lacks, at least partially, access to the evidence that the expert has, is unable to correctly assess the soundness of the reasoning on which the expert bases his conclusion, and does not have access to studies critical to the expert's position (Goldman, 2011; Hardwig, 1985, 1991). Nonetheless, nonexpert might have reasons for believing that the opinion of a given expert is sound, and even might have reasons to believe that this particular expert is more reliable than her opponent (Goldman, 2011). In the latter case, Alvin Goldman (2011) posits that layperson makes an inference about levels of expertise of rival experts. Albeit, I would argue that informed trust in an expert's opinion, in general, can be called inferring to the best expertise, on the grounds that it necessarily includes the stage of comparing a given opinion with the position of other experts.

6. Expert's credibility and reliability

The question of trust in expert opinion can be whittled down to two separate but related issues. The first is the problem of establishing an expert's credibility, and the second is connected with an attempt to enact a level of reliability of his opinion. Whenever we meet with the opinion of a putative expert, regardless of whether it is a public debate, any discussion, or even in a private conversation, our first step should be to establish the given expert's credibility.

To achieve this goal, we should first establish whether the author of the opinion is an expert in the relevant field. There is the crucial distinction between a reputational expert, that is, a person who is perceived as one, and an actual specialist; when the former role is discretionary and may be filled by anyone, even a celebrity, the latter is based on objective premises (Goldman, 2011). Expertise in an irrelevant field can create such a reputational expert too. As I mentioned before, the range of every expertise is

invariably limited to the specific domain, and an expert's opinion that exceeds their area of interest is not much better than that of a layperson. Unfortunately, we tend to effortlessly ascribe authority to someone when they should not have any, so being aware of the limitations of expertise is invaluable. Far-reaching specialization means that it is no longer enough to be a physicist to talk about climate change causes; a more narrow specialization is needed, which in this case is climatology. Climate science is an exceedingly complicated field; only a dedicated specialist can be up to date with the latest research and findings.

Taking the above into account, relying on the opinion of an expert whose area of specialization is adjacent to the proper one may be deceptive or even lead us astray. Undoubtedly, among particle physicists, we will find many familiar with climate science, but their knowledge will always be simplified and limited compared to that of an experienced climatologist. At the same time, we encounter many climate deniers among physicists. It is true in any other discipline; some philosophers, historians, psychologists, sociologists, and even laypeople will be more or less informed, and some will be unquestionably ignorant. However, none of them can match the knowledge of experts in climate science. Therefore the first cue is as follows: if an opinion O is within a subject domain S, expertise of a person E who asserts that O is true (false) should be in S too (Walton et al., 2008; Walton, 1996).

Determining the pertinence of a range of expertise is essential; however, establishing an expert's credibility is not a sufficient condition. There are other cues, which can be supportive in this task. Each expert can boast a history of opinions given, based on which their credibility can be determined; it often involves such issues as absence or presence of frauds, conflict of interest or documented attempts of concealments of such conflicts, plagiarism (Goldman, 2011). It is worth pointing out that not every industry-funded scientist is undeserving of our trust; it depends on the whole social context of their activity, mainly is their opinion is a product of the scientific community, namely, did they attend a conference and publish their paper in a peer-reviewed journal. If this is the case, then we are free to assume that all critical norms and scientific scrutiny are satisfied, and their contribution to the field is as good as any other (Oreskes, 2019). There is a reason why a given expert's social background is among crucial cues of their

credibility. The quality of research that cannot be found anywhere else is precisely the product of various procedures regulating scientists' work. This epistemic quality of research, which cannot be found anywhere else, is the result of the various procedures which formulate the work of scientists. Such quality can only be achieved within a community that meets certain conditions, such as considering and testing many alternative hypotheses, allowing multiple competing points of view, self-criticism, an evidence-based approach to eliminating hypotheses, replication, and modification of conducted research. For example, Fred Singer, a prominent rocket scientist, has been involved in many initiatives sponsored by the tobacco industry, the purpose of which was to cast a shadow of doubt on the scientific evidence linking smoking with lung cancer (Oreskes and Conway, 2010); furthermore, his claims have not been published in any peer-reviewed journal.

On this basis, distrust of Singer's opinion on the causes of climate change is reasonable. His expertise is not pertinent; other experts do not review his views and do not review his view since the so-called merchant of doubt considerably undermines his authority and credibility. Being a merchant of doubt does not ensure that Singer's opinions are dubious (however, it significantly reduces their importance); there is still a chance that his opinion expresses a scientific consensus on climate change. To rule this out, one must compare his words with other researchers' positions and decide if it is consistent with what other expert assert. The risk that we are dealing with a view designed only to spread disinformation is insignificant, on the basis that the greater the expert's agreement on X, the greater the likelihood that the evidence available to humanity supports this particular view. In this particular situation, consistency with other experts' opinions outweighs the unreliable source of information. Even a broken clock is right twice a day, but it is reasonable to trust it only if it is consistent with other clocks. Suppose a given expert opinion is inconsistent or even contradicts the scientific consensus. In that case, it is safer to reject such a view on the basis that the likelihood of the situation where the majority of experts have gone awry and we have met a prodigy presenting a groundbreaking discovery is much lower than the likehood that experts are correct, and alleged prodigy is a fraud. To simplify, let us assume that expert's position in discipline X is true in 51 cases out of 100. That is enough reason for a layperson to

prefer the expert community's coherent view over their guessing, other non-expert opinions, or the opinion of a dubious and lonesome scientist. One may ask why one should ever worry about expert credibility when consensus is much more critical. This issue is crucial as a credible expert is often an excellent source to inquire about whether there is a consensus; it is also much easier to establish the credibility of one expert and focus on the consensus question afterward than to check the entire community's position in the first place.

Unquestionably, trusting in science is always risky; after all, sometimes science makes mistakes, and the position of science is not developed once and for all. New evidence may force it to change and, consequently, the layperson's opinion should be updated. In other words, the dilettante's opinion on issues examined by science should be, "science's position is my position," not because science is the only reliable source of knowledge, but because—as Bertrand Russell notes—when the experts agree on something, the opposite view cannot be regarded as certain (Russell, 2004). Above thought can be expressed by paraphrasing Alvin Plantinga's maxim: "When any belief and science clash, 'tis belief must go to smash" (Plantiga, 2018, 226).[1] This slogan can be developed as follows: "Where it conflicts with common sense, religion, and tradition, science should be regarded as authoritative for education and public policy as well as objective inquiry; and scientific knowledge is even relevant to moral and political deliberation" (Ladyman, 2018, 106). What if science clashes with another science.

7. Disagreement among peers

The proposed heuristics offer some guidance when there are two or even more experts with rival opinions. Whenever we face contradictory testimonies of experts within a pertinent domain of expertise and whose history is untainted by suspicious activity, our last resort is the very question about scientific consensus and its relation to these testimonies. Our trust should

[1] In his article, Alvin Plantinga focuses on the conflict between scientism and religion; hence the maxim he quoted was, "When faith and science clash, 'tis faith must go to smash." I have taken the liberty of generalizing it to all beliefs.

be given proportionally to the support of the opinion given by the community of experts, or as some social epistemologists call this approach, we should "use the numbers" (Coady, 2006; Goldman, 2011).

Consider the dispute over X; scientists have proposed three solutions: A, B and C. Each group supports their position with some scientific evidence, such as completed experiments, the proper amount of measurements, or other analyzed data types. However, experts disagree on quality of those, and as a result, 40% support solution A, 33% support B and 27% think that C is the best answer. The layperson is universally unable to judge the whole body of evidence behind any of these options, but can judge the experts' credibility. All groups consist of professionals with a similar level of trustworthiness, there is nothing suspicious in their previous activities, they have published the whole body of evidence in peer-review journals. There is nothing else for a layperson to do but to assign these positions no more significant degree of belief than professionals' amount of support. The conclusion is that none of the proposals put forward by scientists can be considered as the science position.

There is no single answer to what percentage of a given scientific community must agree to describe theirs as the position of science and treat it as a reliable stance on some issue. It all depends on the particular issue and the context in which it is being considered. When we wonder whether to use a homeopathic remedy, we only need the qualitative information that the vast majority of medical specialists consider such therapies ineffective. For a politician who must decide whether to regulate the legality of such treatments in the state, information about the "vast majority" will not be enough. Determining whether 61% or 91% is behind the term "vast majority" is of great importance in this case. However, knowing that there is no "vast majority" or there is no majority among experts at all will be always compelling, and should be treated as a serious reason not to prefer any of the positions.

If asked today, no one will have a problem with answering whether the iguanodon was a bipedal animal, because the position of paleontology in this matter is unambiguous. It was different in the first half of the 19th century, when paleontology was a fledgling discipline, and the incomplete reptile skeleton had just been discovered. Two paleontology pioneers argued

about the dinosaur's posture. The first, Sir Richard Owen, considered the animal to have been four-legged, and the second, Gideon Mantell, two-legged. The lack of a complete skeleton allowed some freedom in how to reconstruct the shape of the dinosaur. It was not until discovering other fossils that this dispute was resolved in favor of Mantell's position. Until then, a layperson could do nothing but suspend their judgement, on account of the lack of agreement between experts.

There are possible scenarios where the problem under consideration is within a domain where experts are difficult to identify, or the given field is so straightforward that no expert opinion is required. There are a plethora of different fields in which the existence of cognitive experts is at least questionable. There are undoubtedly authorities in such domains, although not every authority, however influential, is based on cognitive expertise. An example of the fields I am referring to may be most areas of the humanities, theological considerations, or even religions. There will undoubtedly be some expertise in these areas related to their history or doctrine content. The existence of such established consensus is not under contention here, although it is crucial to make a distinction between agreement on what Plato's, Aristotle's, or Kant's concept of metaphysics was about and agreement on the fundamental nature of reality itself. The former is a matter for the history of philosophy; the latter is a genuine metaphysical issue. As far as the history of philosophy is concerned, there are reliable experts within this domain, just as there are reliable experts in physics's history. It is worth to emphasize that from the perspective presented here, it is of paramount importance whether there is a consensus on a given issue or not; accordingly, the discussion about the existence of experts can be relegated to the background as an attempt to establish whether there is a consensus or not plays a decisive role. Therefore, a question about metaphysics should be stated as follows: is there any metaphysical issue for which most experts have established a solution? The answer to this question is negative. Not a single problem has been solved in terms of which most metaphysicians agree. Plato's metaphysical system competes with Aristotle's system and every other set of metaphysical beliefs. Therefore, as in the Mantell vs. Owen case, layperson could do nothing but suspend their judgement about metaphysical issues.

Similarly, not only is Islam a holistic alternative to Christianity, but so is any other religion. Textbooks in various fields of science are a good illustration of this point. Books explaining the principles of thermodynamics or the theory of evolution refer to the current state of human knowledge while at the same time, they inform about issues on which there is consensus between experts in a given discipline. Such texts involve many simplifications, which does not change the fact that they contain a set of findings, i.e., statements considered true. There are no textbooks of metaphysics that contain or present the current state of knowledge about the nature of reality because there is not even the slightest consensus on this matter. There is no question of textbooks in the case of religion because there are holy books that are expositions of a specific faith, and there are religious studies that describe various doctrines. The reason for this difference is that the sciences have established certain things. The position of science regarding the number of planets in the solar system is unambiguous. Thanks to the work of astronomers, we know that there are eight of them. There is neither a philosophical nor religious position on the number of existing gods; it is impossible to designate even the smallest number of deities common to all known faiths. Each philosophical and religious system proposes a pantheon that is unique to it, filled with a different number of various gods. Even if some religions postulate one god's existence, they attribute different properties to it and suggest different methods of communicating with it, thus explicitly excluding any similarity between them. In such a case, the extension of the term "expert" to include the authors of metaphysical concepts, founders of religions, theologians are acceptable under the assumption that expertise is gradable. Such disciplines as metaphysics, epistemology, or theology can be treated as fields in the pre-paradigmatic phase, to use Thomas Kuhn's term, as these areas are always torn by disputes over fundamental issues none of them has an established consensus. Therefore, experts in these fields are experts whose reliability is limited. However, this is a consequence of the application of a general heuristic: if a particular field is lacking even the slightest consensus, then before someone decides to trust one of the concepts presented in those domains, they should indicate why we ought to prefer this and not another position. Otherwise, acceptance of any particular position in the unresolved dispute gives rise to the risk of making an error,

proportional to the aggregate percentage of support for other viewpoints. It is irrelevant if this dispute is within physics, biology, gender studies, or philosophy.

Sooner or later, we will come across issues that will be difficult to be assigned unambiguously to a specific discipline. What kind of expertise is needed to evaluate a given political decision or the overall reform undertaken by a government? Is the opinion of a political science specialist enough, or on the contrary, is a consultation with an economist needed, or is it both? Why not ask a sociologist too or a professor of law. Cases such as these are beyond the comprehension of a single domain; therefore, it is difficult to name the pertinent expertise. It is reasonable to seek advice from an expert within a field of expertise related to the problem under consideration and check for any common ground between their opinions. For example, when we encounter an immense number of negative reviews of a given political reform, even if these opinions differ in magnitude, their common aspect is their negative nature. In such a situation, rejecting any positive review is a way to reduce the risk of adopting an ill-founded view or even a thoroughly inadequate verdict.

Before everything else, there are matters of subjects where no cognitive expertise is needed, besides the opinion of an involved person or a group of engaged people. There is no justification for scientific approach to establishing how to hold a woodcutter's ax; moreover, any experts other than the woodcutter alone are unnecessary. To convene an expert committee to determine the contents of a given refrigerator or someone's pocket is also beside the point and even ludicrous. There is no community of experts capable of telling a father of five which of his children should he kiss first after supper. These are only a few examples, but there are a plethora of different issues, and even domains, where scientific expertise is redundant, and the testimony, intuition, common knowledge, hunch or a guess of a single person is a good source of opinion, and a fair basis for making a decision.

8. Scientophilia

Inspired by the term *Biophilia*, the love of all living things coined by Edward Wilson (1984), I would like to propose a name for the heuristic

presented above: scientophilia—the love of science. Love of science is motivated by the fact that science provides knowledge of the best possible degree of justification and manifests itself in an established consensus among credible experts. I would venture to put forward the thesis that most of us very often behave as a scientophile.

Until 1992, Pluto was considered a planet, but observations made at that time and in the following years enriched our knowledge with new information, which precluded us from calling this object a planet any longer. It turned out that Pluto has a smaller mass than the rest of the bodies coorbiting it, which is a breach of one of the necessary conditions for being a planet. In 2006, after several years of disputes, scientists developed the position that Pluto is a different type of celestial body than previously thought, namely a dwarf planet. The vast majority of us behaved in this matter like quintessential scientophiles—overnight, we stopped listing Pluto among the planets, thus rejecting the view that there are nine of them in the solar system. Insisting on the opposite position would have been unreasonable in this situation. Currently, the whole world is struggling with the severe problem of the COVID-19 pandemic, and most of us, although unfortunately not all of us, try to follow the recommendations of scientists. We cannot independently check whether we are sick, predict how the virus will spread, determine what behaviors are safe or whether animals can infect us. We are condemned to expert opinions, and we trust them because those scientists work in organizations that guarantee their employees' reliability.

In conclusion, the main guideline of scientophilia can be described as inference to the most reliable and attainable expert's opinion. This heuristic name indicates a love of science because looking for scientific consensus is advantageous for establishing a well-informed opinion for a layperson interested in a particular issue. In science, a consensus is not achieved by agreement but by examining evidence supporting different positions. The scientific community comprises groups of qualified experts using a variety of procedures to find the best explanations and theories to explain the evidence they collect. They are involved in such activities as critical discussions, gather, analyze, and evaluate various data and publish their research results in peer-reviewed journals. When the evidence starts to tip the balance to the side of some hypothesis acutely, consensus arises. Therefore, if such a

community of cognitive experts has established a consensus on a particular issue, there is no better cue for a layperson to believe that the subject of a consensus is the most reliable position in that matter. Indeed, a consensus is not always essential, as it is frequently redundant and even impossible to achieve in various subject matters. However, if there is an established consensus in a particular field in which we are interested, adopting an opinion contrary to the position of science is associated with a high risk of adopting a view that turns out to be false, and sometimes even harmful to our health or finances.

Scientophilia has some inconvenient consequences, as it entails a change of belief to reflect changes in science. Contrary to appearances, consensus-based opinions are far from perfect and can change, as its foundation does. The view that there are eight planets in the solar system is applied because of the specific definition of the term "planet", based on the current information about our planetary system, which, in turn, is influenced by the sensitivity of modern instruments used to observe space. Changing any of these elements will affect our knowledge of the solar system. A person following this heuristic in 1991 would have thought that it was quite likely that people were causing a sudden increase in temperatures, but there was no certainty, as there were a considerable group of credible experts who disagreed with the others. In 2019, however, things had changed, as there is almost 100% consensus on what causes climate change; therefore, someone would say that we have such certainty. If in 2034, climatologists agreed that they were wrong and it was not humans that caused the temperature rise, there would be no other choice but to accept the position of climate science.

9. What is scientism?

Scientism most often refers to a specific set of philosophical views on the relationship between science and other disciplines. As Rik Peels (2018, 29) observes, almost every type of scientism can be reduced to a set of statements about "the relation that should obtain between the natural sciences on the one hand and something else—another academic discipline or another realm of reality—on the other". The above characteristic is also how

scientism will be understood in this paper. As a side note, it should be noted that a scientist, by default, associates the term "science" either with physics alone or with the natural sciences, which is a relatively narrow meaning of the word.

One of the positions often associated with scientism is the view described by some authors as "scientific expansionism" (Stenmark, 2018) or "scientific imperialism" (Ladyman, 2018). According to this belief, the boundaries of science are far beyond what we think. Usually, this means that science can answer a much larger number of questions; in particular, it can answer questions that we have not associated with scientific research so far (Stenmark, 2018). The above remarks typically concern problems in the field of law, literature, and politics, as well as philosophy, in its broader meaning (Haack, 2012). Such scientism may vary in strength—its most extreme version refuses to acknowledge questions that science cannot answer. When making claims about our knowledge's current state, scientific imperialism—in its extreme version—is trivially false, and we will probably not find a supporter of such a view. We know that there is a wide range of questions that none of the sciences can answer , from those that each of us faces every day ("Should I drink coffee or tea?") to the more complicated ("What taxes should we introduce in our country?") (Haack, 2012). It does not mean that scientific issues do not play any role in social matters.

On the contrary, its function is difficult to overestimate; e.g., medicine does not tell us whether vaccinations should be mandatory, whether the refusal of a vaccine should be punished, and if so how, but it does inform us about the benefits and disadvantages of vaccination so we can make better decisions thanks to this knowledge. Assuming that scientific imperialism does not make claims about the present state of science but about the future, there is no reason to reject or accept such a position. It is also unclear what would result from the adoption of such a view.

Let us assume that in the distant future, it will turn out that physics or the natural sciences will be able to indicate, from the set of all pressing questions, those that have been wrongly posed and answer the rest. Such a scenario in no way justifies the view that now philosophy, for example, should be done on the model of physics or that we should give it up completely. Instead, we should press on physicists to speed up their work.

We can give up the humanities study only when physics replaces it, not when we think that it is possible. The extreme version of scientific imperialism may appear in a weaker, i.e., local, version. Such a version would occur if someone found that natural science has absorbed a set of issues specific to discipline X. An example of such a position may be the view that metaphysical issues are currently being investigated by cosmologists, making philosophers' attempts to resolve these problems superfluous. Local imperialism is the most challenging position to assess because it collects various concepts, each of which deserves a separate analysis. Weak versions of scientific imperialism do not seem to be particularly controversial. No one will deny that many scientific disciplines have emerged from philosophy, so at least some philosophical questions have been answered scientifically after undergoing appropriate modifications. It is even more difficult to reject the above view when we use the term "science" in the broad sense proposed earlier. The weak version of scientific imperialism, which says that science may or may not expand its borders in the future, expresses a belief in scientific progress; hence an excellent rationale can be found.

The imperialist nature of scientism can be implemented in many ways. The first worthy of discussion is the reductionist version of scientism, or internal scientism, as Stenmark (2018) calls it. In proposing a specifically interpreted "scientization" of disciplines outside the natural sciences field, this view develops the idea behind scientific imperialism. Usually, the process of scientization of a given discipline comes down to its complete reduction to a specific science in the strict sense, e.g., to physics, biology, or chemistry. An example of this is the famous sociobiology project of Edward O. Wilson (1975). This type of scientization can be targeted at a specific discipline or all social sciences and humanities. Internal scientism is a distinctive position because it cannot be analyzed in isolation from a specific project of "scientization." Such a discussion would require high competence in all areas involved in the proposed process, and as such, it goes far beyond the scope of this paper.

The standpoint which Stenmark (2018) refers to as epistemic scientism can be regarded as different from the above understanding of scientism. According to Stenmark (2018), some scientists, philosophers, and thinkers (Rosenberg, 2011; Russell, 1978; Sellars, 1963) can be associated with the

claim that "The only kind of genuine knowledge we can have is the one provided by the sciences" (Stenmark, 2018, 63), or even with the more extreme position that "We are rationally entitled to believe only what is scientifically justified" (Stenmark, 2018, 65). There are many possible variations of this notion, which means that its postulates can take different shapes, depending on their author (Boghossian, 2006; Kitcher, 2008; Rosenberg, 2011). For simplicity, I assume that their common denominator is one of the two theses cited by Stenmark. If both of these statements are treated literally, then finding countless counter-examples for them turns out to be a straightforward task. I know that I have two hands, I know that chess pawns attack only diagonally, I know that I have never been on the moon, I know that bachelors have no wives, I know that I have a mobile phone in my pocket, and I know all this without any help from the natural sciences. Any research methods and instruments used in natural sciences are unnecessary in determining the above facts. Nobody observes a bus stop in different weather conditions to determine the bus schedule; after all, it is enough to check the timetable. Examples of non-scientific knowledge, or Moorean truths, as Rene van Woudenberg (2011; 2018) calls them, can be multitudinous because the amount of knowledge sources other than science is staggering. Thus, when a scientist claims that the only source of knowledge about the world is physics/the natural sciences, they should explain their exclusion of the collection's Moorean truths. The easiest way to get out of this situation is for the scientist to admit to using a very narrow definition of knowledge that deals only with what scientistic knowledge is. In such a situation, its exact content and its consequences should be considered.

What are the consequences of the fact that my knowledge of chess rules is not scientistic? Would the non-scientistic character of a police officer's knowledge of a suspect's guilt be a valid reason to abort the arrest? What about a lumberjack's knowledge of the correct way to hold an ax securely? Human knowledge, like science and scientism, is a vast and blurry concept. Nothing prevents one from cutting out some of its fragments and comparing their properties with others, which is advisable, if only for cognitive reasons. Moorean truths differ in some respects from the knowledge of engineers building solar sails for space vehicles, and these differ from the knowledge

of logicians studying the relationships between various formal systems. To understand the rich world of human knowledge, it is undoubtedly necessary to distinguish its various manifestations. It does not change the fact that exclusive claim that only certain areas of human thought constitute knowledge requires precise clarification of non-knowledge fields. Depending on how one answers the question of the status of other alleged varieties of knowledge, epistemic scientism may turn out to be a false and absurd position or not at all as controversial as it is usually painted.

Another variation on scientism worth mentioning is the ontological version. Again following Stenmark, it can be said in simple terms that this type of scientism can be reduced to the thesis that "[t]he only things that exist are the ones that the sciences can discover" (Stenmark, 2018, 68), or in the words of Carl Sagan: "the Cosmos is all that is or ever was or ever will be" (Sagan, 2013, 8). Scientism, which claims that the entire world is limited to physical entities, is close to some naturalism varieties. Ontological scientism inherits from naturalism all the problems typical of this kind of position, i.e., problems with such issues as the existence of norms, works of art, laws of nature, or logical laws. Logically, this kind of scientism seems to be the strongest position since it entails all the other varieties mentioned above. Accepting that the world is limited to entities described by the natural sciences immediately imposes the adoption of the view that only the sciences provide knowledge of reality, the consequence of which is that all human forms of cognition should be either reduced to them or conducted like them.

In summary, scientism, like science, is heterogeneous. The examples mentioned above of differences in understanding the concept of "scientism" do not exhaust the rich semantic field of this term. A particular case of scientism needs not to be limited to epistemological or ontological versions; the above varieties of scientism can often be combined, which is often the case. Most instances of extreme scientism, that is, one which claims that "science is the only..." are very easily dismissed as absurd or even merely false. It is different in the more moderate versions, which are much more challenging to evaluate without the theoretical context in which they occur. In particular, it is impossible to evaluate them without comparing them to competing positions.

10. The missing link in the scientism debate

Discussions about scientism understood as a synthesis of many philosophical positions dominate the literature devoted to this issue. There is also plenty of polemic with particular authors who admit to being scientists or are accused of such sympathies. It is rare for commentators to analyze alternatives to scientism, and without this, it is impossible to evaluate any position fully. The search for antiscientism can be carried out in two ways. The first is to extract antiscientism from the writings of scientism's critics. The second is to determine its shape based on scientism's presentation, for which it is to be an alternative. Every view, theory, or even a single statement can be criticized from a neutral position; therefore, not every criticism of scientism will contain an alternative to this view.

Furthermore, even if the criticism is not neutral, it does not have to directly express an antiscientistic position; a reconstruction of such a position will be required. The second approach has a significant advantage because it allows one to build a theoretical framework for later attempts to extract antiscientistic positions from specific texts. Thus, this is the one we will start with.

The types of scientism cited earlier provide a good starting point for constructing possible alternatives to this view. Since scientism is associated with scientific imperialism, antiscientistic positions will oppose it. The disagreement with the view that the boundaries of science are much further than we think can be expressed with the help of many different statements that form the basis for different types of antiscientism. One may argue that the natural sciences have now reached their end and that we will not learn anything new thanks to them. In particular, they will never enter the realms of law or philosophy. As with scientistic imperialism, there is no reason to reject or accept such antiscientistic imperialism. It is impossible to decide where and when the development of the natural sciences will stop.

Much more radical opposition to scientific imperialism is also possible. According to extreme imperialistic antiscientism, the natural sciences have either long exceeded their powers or, indeed, have never had them because they can be wholly reduced to social sciences (e.g., sociology) or other fields of culture (such as philosophy or poetry) and the issues they study are just

social constructs. Alternatively, no discourse is distinguished, so there is no question of crossing boundaries. Supporters of social constructivism, feminist philosophy of science, or methodological anarchism would probably agree with such claims (Burr, 2003; Feyerabend, 1993; Harding, 1991). Internal antiscientism has automatically emerged from the above statements, one which in its most extreme version will proclaim the reduction of natural sciences to a discipline belonging to the social sciences, possibly to philosophy, or even religion or theology. Opponents of internal scientism do not have to be so radical—they can settle for a much weaker position and proclaim the view that specific disciplines cannot be reduced to natural sciences. This impossibility may be absolute or limited to the current state of knowledge.

To negate the extreme version of epistemic scientism is enough to agree that Moorean truths belong to the set of knowledge and deny that its only source is the natural sciences. Of course, epistemic antiscientism can take an extreme form, not so much pointing to sources of knowledge other than scientific ones, but limiting human cognition to only one sphere related to intuition, mystical experience, or some form of philosophical insight into the essence of things, for example. Thus, epistemic antiscientism would exclude the natural sciences for not being a credible source of knowledge or put them below the alternative of its choice. The natural representatives of ontological antiscientism are various religions and related metaphysical assumptions, but these are usually an extension of the ontology provided by the natural sciences, so antiscientism based on them will be moderate. Other examples of moderate versions of this notion can be provided by various philosophical realisms, i.e., positions that postulate mathematical entities' autonomous existence, moral norms. The extreme version of ontological antiscientism can be found in such philosophical positions as Platonism, recognizing the world of ideas as the only actual reality, or in social constructivism—which I mentioned above—on the assumption that it treats the entirety of reality, including the world of nature, as a social construct.

It is time to look at the actual uses of the term "scientism." Philosophical texts on scientism can be divided into two groups, the first of which comprises debates on the nature of scientism, in which the authors consider how to define this term loosely, how to distinguish its various varieties. The

second group will contain all polemics, containing critiques and defenses written by supporters and opponents of its different varieties. In the latter case, antiscientism will assume different variations on the types outlined above. In addition to theoretical considerations, there are texts in which the word "scientism" functions as an accusation or an epithet used to discredit an opponent's position. Discussions in which such an allegation is made usually concern the conflict between science and some other field, most often philosophy and religion, and the dispute concerning the legitimacy of pursuing the latter. It is not only supporters of philosophy and religion who use this term in this way. In 2015, a public forum entitled "Scientism in the Age of Obama" was held in the United States, with the primary goal of agitating against the various science-based elements of the American president's political program at that time (Pigliucci, 2017).

Another example is an article defending homeopathy against mainstream science. According to this article, mainstream science does not allow homeopathic therapies to be treated as a science because, being possessed by a scientistic ideology, it cannot see the advantages of homeopathic therapies (Ledermann, 2003). Rupert Sheldrake (2012) uses the term in a similar way when he writes about the scientific worldview's followers. This term in the above cases not only serves as an accusation, but it also forms the basis for an appreciation of the defended discipline, or at least elevating it to the scientific level, thereby dismissing any criticism as unfounded.

Assuming that scientism and antiscientism are opposite positions in the dispute over the nature of the natural sciences' relationship and some other field or aspect of reality, their final shape will depend on how this relationship is seen. Using any of these terms on its own does not make sense because a given statement may be considered scientistic by one opponent and antiscientistic by another. One may be convinced that physics will someday displace all metaphysical inquiries like it has displaced Aristotelian physics while claiming that ethics will remain out of its reach. Such a person in a dispute with a supporter of extreme scientism, who claims that physics will also absorb ethics, will take an antiscientistic position. However, to a proponent of the thesis that metaphysics will never succumb to studying the natural sciences, they will appear to be taking a scientistic position. That is why it is so important to define this relationship discussed by

scientism/antiscientism; it is only by defining it that we will identify a proposed position as supporting or opposing.

11. Scientophilia and scientism

Scientophilia, at first glance, resembles a peculiar version of scientism, probably an epistemic one, as it is challenging to identify the single point where a proponent of such a view would disagree with a scientophile. However, if we take a closer look at both concepts, there should be a few crucial differences, although the level of dissimilarity between these two depends on the chosen variation of scientism itself.

The first important distinction between scientophilia and epistemic scientism concerns that scientophilia is not limited to scientific methods or even scientific (in a narrow sense) knowledge. Scientophilia is interested in specific knowledge-forming procedures, and science appears to be an excellent example of its implementation. A scientophile accepts the reliability of any discipline practiced by a community of experts who evaluate each other, who have developed intersubjective methods of evaluating evidence, indulge in critical discussions, and submit their works to journals with peer-review procedures. Thus, science is in the highest place; nevertheless, other levels of expertise are acceptable when scientific expertise is redundant or not attainable. From a scientophilia perspective, disputes over "whether discipline X is scientific or not" are superfluous as long as X's purported experts can reach an evidence-based consensus. However, if a consensus has not been attained, there is no reason to adopt either side's position; it makes no difference if this discipline is physics, philosophy, history of jazz, or religion. Scientophilia also places Moorean truths among consensus-based knowledge, and as far as these are concerned, each of us is sufficiently competent to represent, in specific circumstances, a credible and reliable level of expertise. Therefore in such cases as whether a given person has hand, or how to hold woodcutter's axe there is a consensus among peers. Nonetheless, if any particular epistemic version of scientism can adopt Moorean truths and other non-scientific types of knowledge (law, for example) as a reliable source in cases where scientific expertise is not attainable, this

distinction will begin to fade away; otherwise, scientophilia will appear closer to epistemic antiscientism.

The second difference is inextricably linked to the essential characteristic of scientophilia, where epistemic scientism is a proper philosophical position, scientophilia is instead a decision-making strategy. Scientophilia aims at making our decisions about opinions and testimonies more efficient; thus, its sole advice is to trust any purported expert, only if their opinion is consistent with their peers, eventually to accord a given opinion no greater degree of belief than that found in the expert community. While the primary goal of scientophilia is practical, there is no doubt that scientophilia rests on a theoretical foundation, which includes statements shared with epistemic scientism, especially those that place natural sciences on the podium. Scientism is not scientophilia's only ingredient; for example, there is an aspect of scientophilia that attaches great weight to the significance of consensus, and this feature is based preferably on common sense than on scientism itself. That is because common sense, not scientism, implies trust in the coherent testimony of a group of eyewitnesses when we have not been able to experience the event they saw. Scientophilia adjusts this suggestion to scientific considerations; one should trust in the coherent testimony of a group of credible and reliable experts, whenever he has not been able (usually due to lacking sufficient competence), to gather and evaluate evidence accumulated by those experts.

In terms of other scientism variations, scientophilia is usually theoretically indifferent; however, there are possible areas in which conflict may arise. If scientific imperialism is considered, there is no common ground between the robust version of this position and scientophilia, as the latter is relatively silent about science boundaries. When scientific imperialism states that science can solve any problem, scientophilia only advises adopting the position of the most reliable expert if such opinion is available on the issue. Such an approach is nothing more than informed trust based not on substantive but formal cues, such as credibility and compliance with other professionals' opinions. Scientophilia does not determine what issues a scientific consensus is possible on whatnot, but advises recognizing it if it has already been established. However, there is a possible conflict. Since scientophilia allows the existence of unscientific knowledge and even

endorses experts who are not scientists themselves, it can be reconciled with the notion that there are problems that science has not solved, but that they at least have provisional, though unscientific solutions. As a result, the proponent of scientific imperialism will have a hard time accepting scientophilia as a whole. With weaker versions of scientific imperialism, namely those which express belief in scientific progress, scientophilia can coexist without the slightest problem.

Scientophilia and internal scientism are entirely indifferent to each other, there are no points that would cause any conflict between them, but there are no joint statements for them either. There is no contradiction in the fact that the most radical advocate of the "scientization" of all possible disciplines outside the natural sciences domain simultaneously applies the guidelines provided by scientophiles, at least to the very moment when scientization becomes complete. As for ontological scientism, scientophilia takes no position regarding the existence of anything unless it is about experts and the opinions they express; their existence is, of course, presupposed. With the guidance provided by scientophilia, it is impossible to establish whether something exists or not, but whether we should trust the people who postulate the existence of objects belonging to a given category. It follows that scientophilia is potentially open to various ontological positions, even those which contradict ontological scientism. That is the theoretical level; in practice, scientophilia can be challenging to distinguish from ontological scientism as their verdicts will coincide. After all, consensus on matters such as the number of existing gods or the existence of an afterlife is lacking, but there is consensus on objects such as planets, atoms, or genes.

The relation of scientophilia to imperialist antiscientism resembles the relationship between scientophilia and imperialist scientism. If a given alteration of imperialist anitscientism is a moderate one, it allows and endorses equal every existing narration or possible discourse; then it follows that a scientophilia-based approach is also allowed endorsed. Therefore, at least theoretically, it is possible to embrace the guidelines of scientophilia and simultaneously be a moderate imperialist antiscientist partisan. However, among antiscientistic imperialists, some positions seek to distinguish some discourse, for example, the philosophical or the religious, as a more reliable source than a scientific position. Neither of these stances, by nature,

will be compatible with the guidelines provided by scientophilia because sooner or later, the latter will recommend adopting an opinion that turns out to be contrary to that of the distinguished discourse. The relation to antiscientisitc internalism is correspondingly dual; that is, scientophilia is consistent with any moderate version. However, any supporter of a more radical version of this philosophical position will find deference to the best expertise hard to accept. If agreeing that Moorean truth belongs to the set of knowledge is enough for a given concept to be classified as epistemic antiscientism, then scientophilia should be categorized as such. Of course, in a more radical form of epistemic scientism, the partnership between these two will be limited or even impossible. The ontology assumed by a scientophile is indeed liberal. However, when it comes to various ontologies postulated by ontological antiscientisms, their credibility will depend on whether a consensus of reliable experts supports them, and in most cases, they are not. Among philosophers, there are convinced platonists, but there is not even the slightest agreement between peer professionals about the possible contents, capacity, and other qualities of the world of ideas, not to mention lack of agreement on whether such a plane exists in the first place.

If we agree that mild or moderate scientism embraces the following maxim "When any belief and science clash, 'tis belief must go to smash", then scientophilia occurs as a peculiar variant of mild scientism. However, its central thesis is not about sources of knowledge or the existence of various objects but rather sets out a strategy to support the decision-making process. It should be noted that scientophilia does not claim that the natural sciences are the only source of knowledge; apart from physics and chemistry, it respects the achievements of other sciences, including psychology, sociology, economics, history; it is even able to treat a personal opinion or testimony as a reliable source. Therefore, scientophilia can also be adapted by both moderate scientists and mild antiscientists. Its belonging to one or the other concept depends mainly on the nuances contained in the definition of a given stance. Things get even more complicated when we distinguish local alterations of scientophilia, which, contrary to its global counterpart, is limited to a selected group of problems. Consider a declared phenomenologist, who believes that his philosophical method is the most reliable approach, but unfortunately, its cognitive scope is excessively limited to ethical values.

Whenever she ponders any issue outside of the field of ethics, she can behave like an exemplary scientophile and defer to the most reliable expertise.

12. Problems and limitations

Scientophilia is not without flaws, and one of its most essential imperfections is a problem with, for lack of a better term, self-proclaimed experts such as clairvoyants, psychics, or various pseudoscientists. Communities that bring such individuals together, with an appropriate degree of organization, can create convincing imitations of evidence-based consensuses. There are even peer-review journals dedicated to homeopathy, for example, so detection of such well-crafted deceit requires the enrichment of the scientophilia-based approach with more advanced critical thinking tools.

Another problematic issue is linked to the probabilistic character of scientophilia, as this heuristic does not guarantee by any means that the provided inferences are indefeasible. What seems today to be an established consensus may tomorrow turn out to be a rejected theory. Inference to the best expertise is never definitive, as it aims at providing the most success in the long run, not in a particular case. Hence, false negatives are inevitable. If a consistent scientophile had met Albert Einstein before the entire world of physicists had acknowledged him, it would have been reasonable for him to dismiss his theory. The basis of such a decision is that Einstein's theory was more likely to be inconsistent with classical mechanics because he was a dilettante than it was that he was a lonesome and unrecognized genius presenting a groundbreaking thesis. In this striking example, scientophilia's advice leads to a catastrophic mistake, but it will discard as unreliable an unimaginable number of amateurs at the expense of that single genius when applied dozens of times.

An additional issue is associated with the fact that consensus may be elusive and challenging to identify for a person lacking a good experience and understanding of scientific communication. Such compelling indicators of established consensus as textbooks, reports of prominent scientific organizations akin to the IPCC, WHO, or FDA. are not always obtainable for various reasons, but mostly because there are none. That means consensus

is probably, at least at the moment, part of esoteric knowledge and is recognized only among specialists dedicated to a particular domain within which consensus has been established. Other explanations are possible, too; for example, a consensus has not been fully formed yet. Regardless of the causes, the only solution is to seek the advice of a credible expert. Unfortunately, we do not always meet the latter personally; therefore, our opportunities to ask them consensus questions are minimal.

Occasionally scientophilia will leave its followers with a recommendation to suspend judgment. It is inconvenient because many disputes have a practical dimension, which means taking one side involves taking a particular action. In this case, Suspending judgment is practically equivalent to adopting a position because we will act following one side or, following the other, we will do nothing. For example, if there is a dispute on whether vaccines cause autism and there is a position of evidence-based medicine, opposed by a person who states that she had a revelation, in which the angel announced that vaccines cause autism. Even if we decide to suspend judgment, we will take some action, i.e., to vaccinate or not, which is equivalent to adopting one of the positions.

13. Conclusion

Scientophilia is hardly an alternative for various alterations of both scientism and antiscientism in terms of being a philosophical position. However, scientophilia is not without philosophical assumptions, and those can be treated as potential substitutes. We can say that scientophilia supports science because it is an unprecedented phenomenon in our culture. Undeniably, science has many drawbacks: scientists lie, deceive, make mistakes, or even give in to fashions. However, it is precisely the same as the case of philosophers, priests, historians, homeopaths, law professors. The difference is that scientists sometimes manage to expose these lies, deceptions, mistakes, and fashions, thus choosing the best-justified claims. Therefore, certain things are established in science, which is not always the case within nonscientific disciplines, as there are domains as philosophy, for example, which lack consensus. It does not mean that the latter is devoid of any value or that some form of science should replace it.

On the contrary, philosophy is necessary for ethical or political issues because even with science, we have not worked out anything better than what is offered by the multitude of philosophical positions. If there is no established consensus among scientists on such issues as ethical values, the definition of justice, number of existing gods, sense of life, then from the perspective of scientophilia, there is no difference between a scientist opinion, a philosopher's, or a priest's point of view, as neither of their positions is backed up by a significant majority of other purported experts. However, let us say if philosophers could reach a consensus on any of the above matters, and that consensus resulted from a critical debate on the gathered evidence, followers of scientophilia should adopt such a position like any other consensus reached by experts. Scientophilia is a love for science because the scientific consensus is much more common than in other areas and is also easier to recognize.

On the other hand, if we consider scientism or antiscientism as strategies guiding changes of beliefs, of course, insofar as they contain such guidelines, at least as tacit assumptions. Scientophilia may appear to be a compelling rival for the stronger versions of both positions, mostly when we speak of scientism and anti-scientism, which are simultaneously epistemic, ontological, and imperialist. Such radical varieties of these two can lead to undesirable and even harmful consequences, such as the exclusion of various naive or amateur historical and anthropological theories from the area of pseudosciences, or rejection of reliability of various nonscientific specializations (law, for example) on the basis that they are not scientific (in a narrow sense) fields on the one hand, or the legitimation of religious fundamentalism, numerous forms of relativism, the admission of pseudoscience or even various conspiracy theories on the other. It seems that the view of reliable sources of knowledge adopted in imperialist scientism is too narrow and too wide in the case of its antiscientisitic counterpart. Scientophilia avoids these risks. In the mild versions of both scientism and antiscientism, scientophilia can be adopted as an addendum to them, especially when the cues provided by both of these stances are confusing or indecisive.

The idea of putting scientophilia in the scientistic camp rather than among antiscientistic positions may seem reasonable and tempting, especially since in disputes between philosophy, religion, and science, it will

usually advocate taking the latter's side, as quintessential scientism. However, its consensus-based approach makes this problematic because, on numerous occasions, a science position will not be needed, and there is an acceptable possibility that, in some cases, it will be not preferred. In this situation, scientophilia seems to be closer to being the third way.

References

Adams Jonathan, Pendlebury David, Potter Ross, and Szomszor Martin. 2019. *Global Research Report. Multi-authorship and Research Analytics*.

Blackford, Russell. 2017. "The Science and Humanities in a Unity of Knowledge". In M. *Science Unlimited? The Challenges of Scientism*, edited by Maarten Boudry, and Massimo Pigliucci, 11–31. The University of Chicago Press.

Boghossian, Paul. 2006. *Fear of Knowledge*. Oxford University Press.

Burr, Vivien. 2003. *Social Construction* (Second edition). Routledge.

Coady, David. 2006. "When Experts Disagree". *Episteme*, 3 (1–2). https://doi.org/10.3366/epi.2006.3.1-2.68

Cook John, Oreskes Naomi, Doran, Peter T., Anderegg, William R. L., Verheggen Bart, Maibach, Eed W., Carlton, Stuart J., Lewandowsky Stephan, Skuce, Andrew G., Green, Sarah A., Nuccitelli Dana, Jacobs Peter, Richardson Mark, Winkler Bärbel, Painting Rob, and Rice, Ken. 2016. "Consensus on Consensus: A Synthesis of Consensus Estimates on Human-Caused Global Warming". *Environmental Research Letters*, 11(4). https://doi.org/10.1088/1748-9326/11/4/048002

D'Agostino, Fred. 2016. "Disciplines, the Division of Epistemic Labor, and Agency". In *Social Epistemology and Epistemic Agency. Decentralizing Epistemic Agency*, edited by Patrick J. Reider, 91–109. Rowman & Littlefield.

de Ridder, Jeroen. 2018. "Kinds of Knowledge, Limits of Science". In *Scientism. Prospects and Problems*, edited by René van Woudenberg, Jeroen de Ridder, and Rick Peels, 189–219. Oxford University Press.

Dunbar, Robin I. 1996. *The Trouble with Science*. Harvard University Press.

Fanelli, Daniele 2009. "How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-analysis of Survey Data. *PLoS ONE* (Vol. 4, Issue 5). https://doi.org/10.1371/journal.pone.0005738

Feyerabend, Paul 1993. *Against Method*. Verso.

Goldberg, Sanford C. 2016. "A Proposed Research Program for Social Epistemology". In *Social Epistemology and Epistemic Agency. Decentralizing Epistemic Agency*, edited by Patrick J. Reider, 3–21. Rowman & Littlefield.

Goldman, Alvin 1999. *Knowledge in a Social World*. Oxford University Press.

Goldman, Alvin 2011. "Experts: Which Ones Should You Trust?". In *Social Epistemology: Essential Readings*, edited by Alvin I. Goldman and Dennis Whitcomb 109–37. Oxford University Press.

Haack, Susan 2007. *Defending Science within Reason. Between Scientism and Cynicism*. Prometheus Books.

Haack, Susan. 2012. "Six Signs of Scientism". *Logos & Episteme*, *3*(1), 75–95. https://doi.org/10.5840/logos-episteme20123151

Haack, Susan. 2016. *Scientism and its Discontents*. Rounded Globe.

Hansson, Sven O. 2013. "Defining Pseudoscience and Science". In *Philosophy of Pseudoscience. Reconsidering the Demarcation Problem*, edited by Massimo Pigliucci and Maarten Boudry, 61–78. University of Chicago Press. https://doi.org/10.7208/chicago/9780226051826.003.0005

Harding, Susan. 1991. *Whose Science? Whose Knowledge? Thinking from Women's Lives*. Cornell University Press.

Hardwig, John. 1985. Epistemic Dependence. *The Journal of Philosophy*, *82*(7). https://doi.org/10.2307/2026523

Hardwig, John. 1991. "The Role of Trust in Knowledge". *The Journal of Philosophy*, 88: (12). https://doi.org/10.2307/2027007

Kitcher, Philip. 2008. Science, Religion, and Democracy. *Episteme*, *5*(1). https://doi.org/10.3366/e1742360008000208

Lackey, Jennifer. 2011. "Testimony: Acquiring Knowledge from Others". In *Social Epistemology: Essential Readings*. Edited by Alvin I. Goldman and Dennis Whitcomb. 71–92. Oxford University Press.

Ladyman, James. 2018. Scientism with a Humane Face. In *Scientism. Prospects and Problems*, edited by René van Woudenberg, Jeroen de Ridder, and Rick Peels, 105–26. Oxford University Press.

Ledermann, Erich K. 2003. Saving Holistic Homeopathic Medicine from Mechanistic scientism - An urgent need. In *Homeopathy* (Vol. 92, Issue 3). https://doi.org/10.1016/S1475-4916(03)00039-0

Nickles, Thomas. 2013. The Problem of Demarcation. History and Future. In *Philosophy of Pseudoscience. Reconsidering the Demarcation Problem*, edited by Massimo Pigliucci and Maarten Boudry, 101–20. University of Chicago Press.

Oreskes, Naomi. 2019. "Why Trust Science?" In *Why Trust Science?* Princeton University Press. https://doi.org/10.2307/j.ctvfjczxx

Oreskes, Naomi, Conway, Erik M. 2010. *Merchants of Doubt. How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury.

Peels, Rick. 2018. "A Conceptual Map of Scientism". In *Scientism. Prospects and Problems*. Edited by René van Woudenberg, Jeroen de Ridder, and Rick Peels, 27–56. Oxford University Press.

Pigliucci, Massimo. 2013. "The Demarcation Problem. A (Belated) Response to Laudan". In *Philosophy of Pseudoscience. Reconsidering the Demarcation Problem*. Edited by Massimo Pigliucci and Maarten Boudry, 9–27. The University of Chicago Press.

Pigliucci, Massimo. 2017. Scientism and Pseudoscience: In Defense of Demarcation Projects. In *Science Unlimited? The Challenges of Scientism*. Edited by Maarten Boudry and Massimo Pigliucci, 185–203. The University of Chicago Press.

Plantiga, Alvin. 2018. "Scientism. Who Needs It?" In *Scientism. Prospects and Problems*. Edited by René van Woudenberg, Jeroen de Ridder, and Rick Peels, 219–32. Oxford University Press.

Price, Derek de S. 1963. *Little Science, Big Science*. New York Columbia University Press.

Rosenberg, Alex. 2011. *The Atheist`s Guide to Reality*. W.W. Norton & Company.

Russell, Bertrand. 1978. *Why I Am Not a Christian*. Unwin Paperbacks.

Russell, Bertrand 2004. *Sceptical Essays*. Routledge.

Sagan, Carl. 2013. *Cosmos*. The Random House Publishing Group.

Sellars, Wilfrid. 1963. *Science, Perception and Reality*. Routledge & Kegan Paul Ltd.

Sheldrake, Rupert. 2012. *The Science Delusion. Freeing the Spirit of Enquiry*. Coronet.

Simonton, Dean K. 2018. "Hard Science, Soft Science, and Pseudoscience: Implications of Research on the Hierarchy of the Sciences". In *Pseudoscience. The Conspiracy Against Science*, edited by Allison B. Kaufman and James. C. Kaufman, 77–99. The MIT Press.

Stenmark, Mikael. 2018. "Scientism and its Rivals". In *Scientism. Prospects and Problems*. Edited by René van Woudenberg, Jeroen de Ridder, and Rick Peels, 56–82. Oxford University Press.

van Woudenberg, René. 2011. "Truths That Science Cannot Touch". *Philosophia Reformata*, *76*(2), 169–86. https://doi.org/10.1163/22116117-90000515

van Woudenberg, René. 2018. "An Epistemological Critique of Scientism". In *Scientism. Prospects and Problems*, edited by René van Woudenberg, Jeroen de Ridder, and Rick Peels, 166–89. Oxford University Press.

Walton, Douglas. N. 1996. *Argumentation Schemes for Presumptive Reasoning*. L. Erlbaum Associates.

Walton, Douglas, Reed Chris, Macagno Fabrizio. 2008. *Argumentation Schemes*. Cambridge University Press.

Ware Mark, Mabe Michael. 2012. *The STM report. An Overview of Scientific and Scholarly Journal Publishing*. Accessed October 1, 2019. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1008&context=scholcom

Wilson, Edward O. 1975. *Sociobiology. The New Synthesis*. Harvard University Press.

Wilson, Edward O. 1984. *Biophilia*. Harvard University Press.

Organon F 29 (1) 2022: 165–172
https://doi.org/10.31577/orgf.2022.29107
ISSN 2585-7150 (online)
ISSN 1335-0668 (print)

BOOK REVIEW

Jared Warren: *Shadows of Syntax: Revitalizing Logical and Mathematical Conventionalism*
Oxford: Oxford University Press, 2020, xx+385 pages

Jaroslav Peregrin*

I must start this review non-traditionally, with an apology. As the author of the book remarks (p. 120. footnote 47), "Peregrin (2017) ... cites my (2015), but seems to indicate that I reject unrestricted inferentialism, despite the paper actually being an extensive defense of unrestricted inferentialism." This, unfortunately, is true. The relevant note in my text was mutilated during my revision of the text based on the proofreading of a native speaker. *Mea culpa, mea maxima culpa*. However, what Warren now writes in his book makes me think that we might perhaps call it quits. Warren, despite knowing about my work, including my *Inferentialism* book (which he refers to in his book), does not shy away from claiming that he is the only current defender of unrestricted inferentialism.

Part I of the book has two chapters. In the first, Warren distinguishes his version of conventionalism from some other versions, reaching the twin characterizations:

> **Logical conventionalism**: Facts about logical truth, logical falsity, logical necessity and logical validity in any language are fully explained by the linguistic conventions of that language.
>
> **Mathematical conventionalism**: Facts about mathematical truth and falsity in any language are fully explained by the linguistic conventions of that language.

Warren rejects that logical claims either survey how we *de facto* use logical words and the sentences containing them, or directly spell out how they are

* Institute of Philosophy of the Czech Academy of Sciences
Institute of Philosophy of the Czech Academy of Sciences, Jilska 1, 110 00 Prague 1, Czech Republic
peregrin@flu.cas.cz

used *de jure*—viz. linguistic rules. This chapter also interconnects Warren's version of conventionalism with naturalism.

In the following chapter the author explains the conceptual framework within which he intends to operate. One crucial thing he points out is that conventions, as he understands them, are *not* explicit stipulations. This is important to keep in mind, for I suspect that many readers may tend to assume that a prototypical convention has to be an explicit agreement. Also, it is not easy to see what "implicit conventions", which thus move to the center of attention, actually amount to. (And Warren is not ideally clear on this score.)

Further on in the chapter, Warren summarizes his understanding of the concept of inference. First, he claims that the inferential characterization of logical constants necessitates rules of greater complexity than the simple ones consisting of n premises and a conclusion; and he indicates that his approach will make use of bilateralism, based on the primitive attitudes of acceptance and rejection. Then he characterizes inference as a psychological process: acceptance and rejection being the most basic "mental states", with inferring being a process that has to do with upgrading the particular cases of these attitudes; and atop of this is inferential rule-following, which amounts to already a very complicated psychological-cum-behavioral pattern.

In particular, a subject S, according to Warren, follows an inferential rule R iff "S is disposed not to violate R, to enforce R in two directions, to comply with R when having the disposition to form attitudes towards all of R's component sentences and to infer according to R when given the chance of having the disposition to continue to accept the premises." Note that here inferring is explained as a process which exists independently of rules; and the following of the rules of inference is its specific version. Hence, from this viewpoint the rules of inference are regulative rather than constitutive—they come to regulate a pre-existing practice.

Chapter 3 is perhaps the most important in the book; here Warren lays out, clearly and explicitly, the fundamentals of his "unrestricted" inferentialism. His two most basic principles are the following (pp. 56, 58):

> **Logical inferentialism (LI)**:[1] In any language, the meaning of a logical expression is fully determined by (some of) the inference rules according to which the expression is used.

[1] This shortcut is not in the original.

> **Meaning Validity Connection (MVC)**: In any language L, the meaning determining inference rules for a logical expression are automatically valid in L.

These characterize Warren's standpoint in general.[2] But he insists that his inferentialism is *unrestricted* (which makes it, in Warren's own eyes, unique), and this is embodied in the following principle (p. 64):

> **Meanings are Cheap (MAC)**: Any collection of inference rules that can be used for an expression can (in principle) be meaning determining for the expression.

In Chapter 4, Warren explains how his unrestricted inferentialism leads to logical conventionalism. At a general level, this is quite straightforward: if meaning is brought into being by nothing but an inferential pattern, and if any such pattern is capable of creating meaning, then conventionalism is forthcoming.

But then, of course, we are led to the question of plurality of logic, which Warren deals with in Chapter 5. It may seem that according to unrestricted inferentialism, it is not only that meaning is cheap, but also that logics are cheap—perhaps all too cheap. What prevents us from establishing a convention by which we make "The moon is made of cheese" or "Saul Kripke was born before Plato" into logical truths? (We can, for example, add the infamous **tonk** of Prior (1960) to current English and we are done, for then any sentence follows from "1+1=2" by means of pure logic). But here Warren makes a crucially important point: we need not block such a possibility, for it does not exist. His point is that though the sentence "Saul Kripke was born before Plato" will be,

[2] Another principle he poses is **Totality**: "In any language L, if a logical inference involving a logical expression is valid in L, then its validity is fully determined by the automatic validity of the meaning-constituting rules for the expression." But this principle seems to me to be superfluous—despite Warren's arguments to the contrary. (In addition, this principle, as it stands, does not seem to be correct. A rule such as *disjunctive syllogism* involves $\vee$, but its validity is obviously *not* fully determined by the automatic validity of the meaning-constituting rules for $\vee$. Plural is required.) It seems that it follows from (LI) plus two other principles, which seem to me to be a matter of course: 1. An inference (rule) is logical if it involves only logical words essentially (if it is presented as a schematic inference, then it contains only logical expressions). 2. The validity of an inference rule is fully determined by the meaning of those expressions that the corresponding schema contains essentially.

in "Tonglish" (English+**tonk**), a logical truth, there is no reason to think that it will mean the same as the homophonic sentence in English.[3]

In Chapter 6, various topics concerning the epistemology of logic are discussed, and Chapter 7 then deals with a traditional objection to basing logic on conventions, Quine (1936)'s argument against Carnapian conventionalism. This concludes the second part of the book, devoted to logical conventionalism. In the third part, Warren turns his attention to mathematical conventionalism.

In Chapter 8, he considers the possibilities and hindrances of extending logical conventionalism, as scrutinized so far, to mathematics. He cites two specific hurdles to be overcome: the first concerns the existence of mathematical objects (for mathematics is replete with existence claims which appear to be hard-won, while conventionalism appears to be able to make such claims true by fiat), and the second concerns the determinacy of mathematical truth (for we know from Gödelian incompleteness that no inference rules can fix this).

The first of these challenges is picked up in Chapter 9. Warren admits that, indeed, on the conventionalist's construal, bringing objects into existence is easy: the existence of an F is secured once our theory entails $\exists xF(x)$. But contrary to appearances, this does not have to compromise conventionalism. We cannot secure the existence of God by accepting $\exists x\mathbf{God}(x)$. Why? It is the same problem as with adding **tonk** to English to make "Kripke was born before Plato" into a logical truth: we can indeed accept $\exists x\mathbf{God}(x)$, but it will claim that what exists is the kind of entity denoted by **God**, not necessarily God.

The other conundrum of mathematical conventionalism, the determinacy of mathematical truth, is handled in the next chapter. To avoid misunderstanding, it is important to stress that determinacy is *not* supposed to contradict mathematical pluralism. We can have alternative and incompatible mathematical theories (as an inevitable consequence of conventionalism). As the author puts it: "Pluralism concerns alternative linguistic practices, determinacy concerns truth in *our* practice" (p. 241). Given this, the problem here is how to overcome Gödelian incompleteness. And to make a long story short, a mathematical conventionalist, according to Warren, can overcome this by taking two measures: by accepting infinitary inference rules (especially the ω-rule, which makes Peano arithmetic complete) and by accepting the open-endedness of rules (i.e. their

[3] I would say that it will not mean the same; however, as Warren does not tell us what he thinks the meanings of empirical sentences are, I am not sure he can put it like this.

persistence throughout expansions of language, for this makes arithmetic categorical).

The remaining two chapters of the book's third part then deal with a lot of possible objections to mathematical conventionalism. The last, fourth, part of the book consists of two chapters devoted to the historical issues regarding conventionalism and to various further philosophical issues related to the author's standpoint.

Before opening the critical part of my review, I should stress—to avoid misunderstanding—that I find the book deeply interesting, stimulating, and original. Warren clarifies many of the issues surrounding inferentialism and conventionalism, and shows that his unrestricted inferentialism is viable, as well as the kind of conventionalism to which it leads. Some of the solutions to traditional puzzles Warren presents are technically brilliant and philosophically revealing. But despite all this, it seems to me that some questions remain unanswered.

I should explain that I myself adopted a standpoint very close to what Warren calls unrestricted inferentialism many years ago, and have long been wrestling with fine-tuning the conceptual framework which is its natural home. With this background, I think that we must make some crucial conceptual distinctions, not all of which are observed by Warren. Let me mention, very briefly, at least three of them. My explanations why they are crucial will be only cursory; discussing them at length is a matter for another occasion.

1. Non-existence vs. uselessness

Warren, we saw, insists that any kind of inferential pattern institutes a meaning. Thus even the infamous pattern governing **tonk**, pace Prior, furnishes the operator with a meaning. I agree that there is no boundary separating meaning-conferring and meaning-non-conferring patterns. On the other hand, it is clear that not all patterns are alike. Some of them, like the one governing **tonk**, are vicious—they wreck any language of which they become a part. And if we agree that something is a language only if it can serve some non-trivial purposes concerning human communication, then nothing containing **tonk** is a language, and hence there is a legitimate question whether **tonk** should be called a meaningful expression.[4] (MAC) states that meanings are cheap; but granting the status of meaning is also cheap—if nothing substantial follows from it.

[4] The situation is reminiscent of that with the analytic/synthetic boundary: there is no boundary separating analytic and synthetic sentences; yet as a matter of fact

Also there is one more boundary that (MAC) does not mention at all: the boundary between patterns that constitute *logical* constants and those that do not (perhaps they constitute something else, like constants of mathematics). (MAC) says that any collection of inference rules furnishes an expression with a meaning, but does it make it into a *logical* constant? This is hardly possible, for then there would be no room, e.g., for *mathematical* conventionalism. So could it not be the case that **tonk** is meaningful, but not a logical constant?

2. Non-epistemic vs. epistemic construal of truth

What Warren writes about the relationship between inference rules and truth is confusing. After stating the principle (MVC) he continues: "Validity requires necessary truth-preservation in the strongest possible sense." How should we interpret the "require"?

One possibility would be that truth is independent of inference (it is correspondence with reality or something tantamount to this), and then inference could be truth-preserving only if it managed to mimic the relation of truth-preservation, which is independent of it. But this, obviously, would contradict the unrestricted inferentialism Warren cherishes.

There remains another possibility: that truth is derived from inference (perhaps it is correct assertability as Sellars, 1968, has it). Then we can say that inference is truth-preserving in a trivial sense, because truth, by definition, is what is preserved by inference. As far as I can see, this is the only possibility compatible with unrestricted inferentialism. But it is strange that Warren tells us nothing whatsoever about this.

3. Natural vs. artificial languages

There are two kinds of languages, natural ones and artificial ones. From the viewpoint of conventionalism, the two kinds are essentially different: while the former are inevitably based on "implicit" conventions, the latter are typically created in terms of explicit stipulations.

Warren starts the book by formulating the inferential rules he talks about for English, like (e.g. p. 45)

---

we will hardly ever give up sentences like "Bachelors are not married" or "1+1=2", so they do have a status that is specific, though only in the pragmatic sense.

$$\frac{\phi}{\phi \text{ or } \psi}$$

Later he writes

$$\frac{\phi}{\phi \vee \psi,}$$

still calling it "or"-introduction (p. 115). This, of course, is ok. Often, when talking about natural language we allow the logical vocabulary of natural language to be represented by its well-known logical regimentations. However, it is at this point that it becomes extremely important (as I have argued at length elsewhere—see Peregrin, 2020) to distinguish between talking about natural language via the artificial proxies of its expressions and when talking of an artificial language composed of the proxies.

Now Warren, after talking about the way in which inferentialism leads to conventionalism, presents a section "The role of semantic completeness" where we can read, e.g. the following passage (p. 107):

> More formally: If we assume that logical truth is extensionally characterized in a language L, semantically, by $\models$, then the conventionalist account requires a proof relation in L, $\vdash$, spelled out in terms of proofs, using the rules of language, that suffices for capturing everything captured by $\models$. If completeness fails, there will be some set of sentences Γ and a sentence ϕ such that $\Gamma \models \phi$, but $\Gamma \nvdash \phi$. This requirement immediately raises a number of serious concerns about incomplete extensions and incomplete alternatives to classical logic.

This is utterly confusing. What is $\models$? Of course, this symbol is standardly used for the model-theoretically defined relation of logical consequence, but could it be that Warren abruptly switches, without warning, from natural to artificial languages? Or does he think that also natural languages have their "model theories"?

So from my (perhaps nit-picking) viewpoint, Warren still has to face some problems he has not addressed in his book. Despite this, I am grateful to him for tabling so many interesting concerns related to inferentialism, and proposing solutions to most of them.

References

Peregrin, J. 2014. *Inferentialism: Why Rules Matter*. Basingstoke: Palgrave. https://doi.org/10.1057/9781137452962

Peregrin, J. 2017. "Is Inferentialism Circular?" *Analysis* 78 (3): 450–454. https://doi.org/10.1093/analys/anx130

Peregrin, J. 2020. *Philosophy of Logical Systems*. New York: Routledge. https://doi.org/10.4324/9780367808631

Prior, A. N. 1960. "The Runabout Inference-Ticket." *Analysis* 21 (2): 38–39. https://doi.org/10.1093/analys/21.2.38

Quine, W. V. O. 1936. "Truth by Convention." In *Philosophical Essays for A. N. Whitehead*, edited by O. H. Lee, 90–124. New York: Longmans. https://doi.org/10.1017/CBO9781139171519.018

Sellars, W. 1968. *Science and Metaphysics*. London: Routledge. https://doi.org/10.1017/S0031819100009645

Warren, J. 2015. "Talking with Tonkers." *Philosophers' Imprint* 15 (24). http://hdl.handle.net/2027/spo.3521354.0015.024