CROATIAN JOURNAL OF PHILOSOPHY

Vol. XXII · No. 64 · 2022

Articles

Book Review

Croatian Journal of Philosophy
Vol. XXII, No. 64, 2022
https://doi.org/10.52685/cjp.22.64.1
Received: October 19, 2021
Accepted: April 5, 2022

Is Representationalism Committed to Colour Physicalism?

DANIEL MARIO WEGER
Goethe-University Frankfurt am Main, Germany

The circularity problem states that the representationalist about phenomenal consciousness gives a circular explanation if she adopts the classic view about secondary qualities, such as colours, that characterises them as dispositions to produce experiences with a specific phenomenal character. Since colour primitivism faces severe difficulties, it seems that colour physicalism is the only viable option for the representationalist. I will argue that the representationalist is not committed to colour physicalism because she can adopt an anti-realist theory of colour. My diagnosis is that the alleged commitment to colour physicalism rests upon the acceptance of colour realism which is due to the approval of externalist versions of representationalism, such as tracking representationalism. I will argue that the representationalist can deal with the circularity problem by adopting figurative projectivism, which holds that colours are contingently non-instantiated properties that only figure in the representational contents of colour experiences.

Keywords: Representationalism about phenomenal consciousness; secondary qualities; circularity problem; colour physicalism; colour projectivism.

1. *Introduction*

Representationalism about phenomenal consciousness holds that the phenomenal character of an experience can be explained in terms of its representational content. Well-known representationalists such as Dretske (1995), Lycan (1996), and Tye (1995, 2000) prefer *strong representationalism,* which has it that phenomenal character is just one and the same as representational content of a specific sort.[1] The significant advantage of the identity claim is that it comes with an account of what phenomenal consciousness is by its nature, whereas *weak representationalism*, construed as the thesis that phenomenal character supervenes on representational content, fails to provide such an account. My focus in this paper will be on strong representationalism because the problem I will be dealing with only afflicts strong versions of representationalism—but more on this later on.

The starting point of the question I will deal with in this paper is that representationalism[2] faces a significant problem with secondary qualities such as colours.[3] Since representationalism explains the phenomenal character of an experience in terms of its representational content, it is not compatible with the classic dispositionalist view about secondary qualities that construes them as dispositions to produce experiences with a specific phenomenal character in normal observers under standard conditions. This is what I will call *the circularity problem* for representationalism. Hence, it is generally accepted that the representationalist must adhere to a theory about colours that does not characterise colours in terms of the phenomenal character of the experiences they are apt to produce in observers. Since colour primitivism is at odds with what empirical science tells us about colour vision and the surfaces of perceived objects, colour physicalism seems to be the only viable option. Thus, representationalism is commonly held to be committed to colour physicalism, and representationalists have been making considerable efforts to vindicate colour physicalism. Yet, colour physicalism itself faces severe objections and is therefore not undisputed. Being committed to a colour theory widely believed to be false, the representationalist finds herself in a very unpleasant situation.

In this paper, I will examine whether representationalism is committed to colour physicalism in the first place. I will give a negative answer to this question and argue that the representationalist can adopt figurative projectivism instead because representationalism, in general, is compatible with an anti-realist theory of colour. Moreover, I will show that only externalist versions of representationalism, such as tracking representationalism, favoured by Dretske, Lycan, and Tye, need to stick to colour realism, i.e., a view that colours are instantiated in material objects.

[1] The qualification 'of a specific sort' is needed because phenomenal character cannot be held to be identical to representational content *tout court*. Obviously, there are mental states with representational content that lack phenomenal character, e.g., standing states like beliefs and wishes or non-conscious occurrent states like those in subliminal perception or early sensory information processing.

[2] From here, "representationalism" is used to refer only to strong representationalism if not stated otherwise.

[3] Since colours are treated as paradigmatic for secondary qualities, I will restrict my argument to the case of colours. Nevertheless, the points made in this paper carry over to other secondary qualities such as smells, sounds, tastes etc.

I will proceed as follows: In section 2, I will elaborate on the circularity problem to clarify why representationalism is incompatible with a specific construal of secondary qualities like colours. Section 3 will state the problems bestowing colour primitivism and explain why representationalism seems thus committed to colour physicalism. In section 4, I will present the objections against colour physicalism and depict an argument against representationalism that emerges from the preceding considerations. In section 5, I will examine the efforts to save colour physicalism undertaken by representationalists and argue that they fail. In section 6, I will show how the representationalist might resist the commitment to colour physicalism by adopting anti-realism about colours. Moreover, I will explain why the commitment to colour physicalism rests upon the approval of an externalist version of representationalism, such as tracking representationalism. In section 7, I will examine anti-realist theories of colour and argue that figurative projectivism is compatible with representationalism and offers a promising alternative to deal with the circularity problem. Section 8 will clarify what makes figurative projectivism attractive and how some prima facie problems might be attenuated. Finally, I will give a short outlook on the ensuing consequences for the prospects of representationalism in section 9.

2. *The circularity problem for representationalism*

At the heart of what I will deal with in this paper is what I will call the *circularity problem* for representationalism. Michael Tye (1995: 144) depicts the setting as follows:

> On the face of it, colors and other "secondary qualities" (smells, tastes, and sounds, for example) pose a special difficulty for the theory I have been developing. If these qualities are subjective, or defined in part by their phenomenal character, then what it is like to undergo the experiences of such qualities cannot itself be understood in terms of the experiences' representing them. That would create an immediate vicious circle.

This statement suggests that secondary qualities, such as colours, pose a particular threat to representationalism. Moreover, the potential problem facing representationalism is described as a case of circular reasoning. Thus, we need to look at two issues: What is it about secondary qualities that makes them a problem for representationalism? And what is the vicious circle that threatens representationalism?

What does it mean that secondary qualities are "subjective"? To start with, primary qualities like shape, size, and motion are commonly held to be properties that are intrinsic to the objects that possess them and, therefore, observer-independent. This means that these are qualities that objects have irrespective of whether they are possibly perceived or not. Hence, primary qualities can be characterised without appealing to how they might affect observers. In contrast, secondary qualities are usually construed as qualities defined in terms of subjective responses, i.e., how they might affect perceiving subjects. Therefore, what characterises a secondary quality as the quality it is, are the responses this quality is apt to produce in perceivers. To the extent that the quality's nature is thus dependent on what responses it possibly causes in a perceiving subject, it is subjective in a sense.

With this in mind, we can also make sense of the clause that secondary qualities are "defined in part by their phenomenal character" if we assume that the subjective responses they are apt to produce in perceivers are essentially characterised by their phenomenal character. Accordingly, we can conceive of secondary qualities as defined in terms of the phenomenal character of the experiences they are apt to produce in the subjects that perceive them. When applied to colours, we receive the view that colours are characterised by their being disposed to produce visual experiences with a specific phenomenal character. For example, something is red just in case it is apt to produce sensations with a reddish phenomenal character. However, something needs to be added to the present account since the fact that objects might produce different responses in different observers and under different conditions is not compatible with the common-sense intuition that each object possesses only one real colour. This problem is usually fixed by adding the two qualifiers of *normal observers* and *standard conditions* to determine the colour property of a specific object. Thereby, we receive the following characterisation of the property of being red:

> X is red $=_{\text{def}}$ X is disposed to produce experiences with reddish phenomenal character in normal observers under standard conditions.

Such a view amounts to giving a dispositionalist account of colours because it defines colours as dispositions to produce visual experiences with a specific phenomenal character in a particular class of observers under certain conditions.

Now, we can see what problem arises for representationalism. As already indicated by Tye, representationalism faces the threat of running into a vicious circle. This problem is clearly brought out in the following passage of William Lycan (2019):

> [O]ne could not (without circularity) explicate phenomenal greenness in terms of represented real-world public colour and then turn around and construe the latter real physical greenness as a mere disposition to produce sensations of phenomenal greenness, or in any other way that presupposed phenomenal greenness.

On the one hand, representationalism has it that, say, an experience with greenish phenomenal character can be explained in terms of its representing that something is green. On the other hand, the dispositionalist theory of colour states that something is green just in case it is disposed to produce experiences with greenish phenomenal character in normal observers under standard conditions. Combining these two claims obviously gives a circular account because phenomenal character is explained in terms of representational content, and the properties that figure in the representational content of a relevant experience are explained in terms of the phenomenal character of the experiences they are apt to produce. This circularity is fatal, for it undermines the representationalist's aspirations to give an informative account of phenomenal consciousness.

Note that the circularity problem only concerns strong representationalism, which claims that phenomenal character is identical with representational content of a specific sort. In contrast, weak representationalism might adopt the view about secondary qualities presented above because it solely holds that phenomenal character supervenes on representational content. This implies that there can be no change in phenomenal character without a corresponding change in representational content. However, this does not require defining phenomenal character in terms of representational content—as strong representationalism does—and, therefore, no circularity is threatening. Nevertheless, it does not seem to be the right move to advert to weak representationalism in the face of the circularity problem since the notion of supervenience is not an explanatory one but serves as a starting point for further investigation into the relation between phenomenal character and representational content rather than being a substantial theory about phenomenal character.

So, suppose the representationalist wants to avoid the circularity problem. In that case, she must give an account of colours that does not define them in terms of the phenomenal character of the experiences they are apt to produce in perceivers. In short, a dispositionalist theory of colour, as presented here, is not an option for her.[4] To put it straight, the circularity problem brings about the following constraint, (C), for a representationalist theory about phenomenal consciousness:

> (C) If representationalism is true, then colours must not be defined in terms of the phenomenal character of the experiences they are apt to produce in perceivers.

Now, it is time to look at which theories of colour satisfy the requirement stated in (C). This will be the topic of the next section.

3. *Why representationalism seems committed to colour physicalism?*

Up to this point, representationalism has not yet encountered any substantial difficulty. To be precise, the circularity problem only constrains the representationalist's choice regarding the metaphysics of colours. So, let us now look at the options she has. The two most influ-

[4] The very idea of a dispositionalist theory of colour does not entail a characterisation of colours in terms of phenomenal character. It is possible, for example, to define colours as dispositions to appear or look a certain way. However, such versions of dispositionalism face the problem of delivering a circular account (see Boghossian and Velleman (1997) and McGinn (1996); for a response see Byrne and Hilbert (2011)).

ential accounts of colour satisfying (C) are colour primitivism[5] and colour physicalism.[6] Colour primitivism holds that colours are simple and intrinsic properties sui generis, whereas colour physicalism claims that colours can be identified with physical properties of material objects. While primitivism conceives colours as non-analysable, non-relational, and non-reducible properties, physicalism assumes that colours can be reduced to physical properties of some kind.

Let us first consider colour primitivism. It enjoys a lot of prima facie plausibility because it matches our common-sense intuitions about colours. According to colour primitivism, colours just are what they phenomenologically appear to be: qualities that populate the surfaces of the things we visually perceive and are involved in bringing about colour experiences. Although colour primitivists hold that colours correlate with the physical properties of surfaces in some way, they reject that the first can be reduced to the latter. The primary motivation for this view is that the essence of colours is fully revealed in visual experience (Byrne and Hilbert 2007; Johnston 1992).

While this view is intuitively plausible, it faces several serious objections. Here, I will focus on the most pressing ones: First, colour primitivism flies in the face of what science tells us: Our best scientific theories, including those concerned with colour vision, suggest that material objects are not coloured since we can explain how colour experience comes about without having to allude to primitive colour properties. It is sufficient to advert to the physical properties of light, the perceived objects, the perceiver's visual system, and the lighting conditions to comprehensively explain how colour perception works (Gow 2014: 809; Maund 2018; Rubenstein 2018). Thus, properties like those postulated by the primitivist are explanatorily idle. Suppose the primitivist nevertheless holds that material objects instantiate primitive colour properties. In that case, she faces the following dilemma: Either she claims—contra what empirical sciences suggest—that primitive colour properties are involved in the production of colour experiences, or she adopts the view that colours are causally inert epiphenomena. While the first option comes with the cost of embracing causal overdetermination, the second one contradicts the common-sense assumption that the colours of objects are involved in the production of colour experiences (Gow 2014: 809; Hardin 1988: 61; Byrne and Hilbert 2007: 82–85). Therefore, neither option is plausible.

Another forceful objection against colour primitivism has it that the same external world object can look different concerning its colour on different occasions and to different perceivers. This results in the following dilemma: Either we accept that it has more than one colour,

[5] This view is held by Campbell (1997), Hacker (1987) and McGinn (1996), for example.

[6] Well-known defenders of this view are Armstrong (1997) Byrne and Hilbert (1997, 2003), Smart (1997) and Tye (1995, 2000).

or we need some non-arbitrary way to decide which one is “the right” colour of the object. Taking the first option flies in the face of our everyday assumption that objects only have one colour and that they do not change their colour every once in a while. The problem with the second option is that there is no non-arbitrary way to determine the conditions under which the object’s actual colour is revealed (Hardin 1988: 80).

Altogether, these objections suggest that colour primitivism is in bad shape and that the representationalist should not adopt it. And now, we can see how the circularity problem for representationalism, together with the shortcomings of colour primitivism, ultimately leads to the claim that representationalism is committed to colour physicalism: The need to satisfy the constraint stated in (C), which is a result of the circularity problem, combined with the fact that colour primitivism fails leads to the claim that colour physicalism is the only viable option for the representationalist. However, the representationalist might remain unshaken. If she is willing to accept colour physicalism, she still does not face any more profound problems. Only if there were compelling arguments against colour physicalism would the representationalist find herself in an unpleasant situation. But unfortunately, this is precisely the case, as I will show in the next section.

4. *The problems with colour physicalism*

Colour physicalism claims that colours can be reduced to physical properties. One way to accommodate this claim is to hold that colours are identical to spectral reflectances of surfaces. The spectral reflectances of a surface is its disposition to reflect and absorb a certain amount of the incident light at every wavelength of the visible spectrum (Byrne and Hilbert 1997, 2003; Tye 1995, 2000). Yet, this view is susceptible to several objections. First, it is phenomenologically inadequate because colours do not look like surface spectral reflectances in visual perception.[7] If I visually experience a red object, it does not look to me as if the object instantiated such-and-such a spectral reflectance. Instead, it seems that the perceived object instantiates a specific qualitative property at its surface that is also had by ripe tomatoes and fire engines. So, the nature of colours as given in experience is radically different from what colour physicalism tells us (Averill and Hazlett 2011; Campbell 1997; Mendelovici 2018).

Second, colour physicalism cannot account for the alleged truth of claims about similarity relations between colours and their structural features, such as “Orange is more similar to red than to green” or “Red is a unique hue, whereas orange is a binary one” because there is nothing about surface reflectances that renders these statements true (Hardin 1988; Pautz 2006).

[7] A similar objection has been raised against colour dispositionalism in its realist version by McGinn (1996).

Third, the problem of metamers brings major trouble to colour physicalism. The starting point for this objection is the empirical fact that objects with different surface spectral reflectances can look the same for a specific observer under certain lighting conditions. Therefore, colours cannot be identified with spectral reflectances (Hardin 1988). A usual response on behalf of the colour physicalist is to identify colours with sets or disjunctions of spectral reflectances properties (Byrne and Hilbert 1997, 2003). Yet, this proposal comes to nothing because the only thing that keeps together the elements of a specific set is their aptness to look the same or produce visual experiences with the same phenomenal character (Gow 2014: 806).[8] This, however, means that colour physicalism would give up its aspiration to define colours without recourse to possible subjective responses.

Overall, the considerations just presented give us strong enough reason to accept the claim that colour physicalism is quite implausible. Taken together with the assumption that representationalism is committed to colour physicalism, this yields a disastrous conclusion for the representationalist. Now, there are two reactions on behalf of the representationalist. On the one hand, she might refuse the claim that colour physicalism is false and defend it against the abovementioned objections. On the other hand, she might argue that representationalism is not committed to colour physicalism in the first place by either rebutting the circularity problem and its consequences or by showing that representationalism might be combined with another theory of colour. So far, representationalists have usually taken the first route and tried to vindicate colour physicalism, for example, Dretske (1995) and Tye (1995, 2000). In the next section, I will examine their efforts to save colour physicalism and show that they fail.

5. *Defending colour physicalism*

In *Naturalizing the Mind*, Dretske (1995: 88–93) defends the view that colours are objective properties. Though he does not fully embrace the idea that colours are identical to spectral reflectances of surfaces, he acknowledges that the latter might play a significant role in characterising what kind of objective properties colours are. Nevertheless, his position should be considered a version of colour physicalism because he holds that the objective properties with which colours are identical can be characterised in broadly physical terms. In this context, he addresses the problem of metamers and appeals to the fact that the visual system of humans was naturally selected under some specific circumstances because it enabled humans to identify the colours of objects under these very circumstances and, thus, helped them to flourish and survive. According to Dretske, metamerism results when the

[8] This point is even acknowledged by Dretske (1995: 89–90) who otherwise defends colour physicalism. See also the section 5 of this paper.

visual system of humans operates under conditions for which it was not originally selected. So, for him, metamerism is a case of perceiving colours under conditions that deviate from selection conditions and, therefore, misrepresentation is just what we should expect. But this, Dretske continues, in no way implies that colour experiences do not represent objective properties.

This line of response is not appropriate because the core of the problem of metamers is that two or more surfaces with different spectral reflectances look the same under some specific lighting conditions. In contrast, Dretske considers cases of objects with different objective properties—spectral reflectances, for example—looking similar under different conditions. This, however, misses the point stated by the problem of metamers. Moreover, there is no reason to think that metamerism could not occur under what Dretske calls selection conditions. To defend his position, Dretske would need to show that metamerism can only happen when selection conditions do not obtain. His claim that metamerism always involves illusion is extremely hard to swallow (Shrock 2017: 141–142). Finally, the assumption that there must be some objective property represented in colour experience seems to be nothing more than wishful thinking since it is doubtful that colour vision even has the biological function of detecting physical properties of objects. As Ross (2000: 123–124) remarks, it is much more adequate to account for colour vision's biological function in ecological terms.

Let us turn to another well-known representationalist who tackles the circularity problem by trying to vindicate colour physicalism. In *Ten Problems of Consciousness*, Tye (1995: 146–148) defends a view that identifies colours with ordered triples of spectral reflectances, each of which covers a specific band of wavelengths corresponding to the wavelength ranges to which the three different types of cones in the human eye are each sensitive to. Given this proposal, Tye wants to account for the problem of metamers by holding that metamers have similar triples of spectral reflectances. This allows for objects with different spectral reflectances to have the same colour as long as their spectral reflectance properties are similar enough concerning the relevant bands of wavelengths. Moreover, Tye thinks that his proposal can also account for the similarity relations between colours because the relations between the triples of spectral reflectance mirror the structure of the hue circle.

While this is an interesting proposal that could be empirically assessed, it is purely speculative that metamers have similar triples of spectral reflectance. Anyway, I think that even empirical evidence will not do it. This response is unsatisfying because it provides an anthropocentric account as far as the selection of the relevant wavelength bands solely rests upon a contingent fact about the visual systems of humans. So, whether the proposal can eventually be empirically corroborated or not, it fails as an objective theory of colour because it defines colours relative to the structure of the visual system of a specific species. Yet, it is conceivable that the visual system of another species has a different number of cones or that its cones respond to different wavelength bands, for example. Nevertheless, these animals might have experiences with bluish, reddish, and yellowish phenomenal character. How should we then decide what triples or n-tuples of spectral reflectance colours are? On the one hand, it seems arbitrary to define colours relative to the visual system of a specific species. And on the other hand, taking all possible combinations of n-tuples of spectral reflectance into account delivers a disjunctive characterisation of colours that is eclectic rather than objective.

In *Consciousness, Colour, and Content,* Tye offers another proposal that incorporates insights from the opponent-process theory to deal with the problem of metamerism (2000: 159–161) and the fact that colour physicalism cannot account for the alleged truth of statements about similarity relations between colour (2000: 162–165). This proposal is prima facie plausible because it provides a characterisation of colours that is insofar objective and observer-independent as activity patterns in the opponent process channels are objectively discernible and quantifiable. However, a severe problem with this proposal becomes apparent upon closer inspection. Why would we accept the claim that some colour is identical to a set of conditions or properties that cause a specific activity pattern in opponent-process channels? I suspect that the inclination to accept such a claim rests on the assumption that some particular activity pattern in opponent-process channels produces experiences with a specific phenomenal character. But this results in a dilemma for Tye: Either he accepts this assumption and gives an account that indirectly characterises colours in terms of the phenomenal character of the experiences they are apt to bring about. Or he rejects this assumption and claims that there is no relevant connection between activity in opponent-process channels and the phenomenal character of colour experience. But in the latter case, it is utterly mysterious why activity in opponent-process channels should then be an appropriate candidate for the characterisation of colours at all or a better one than any other neuronal activity that also bears no relevant connection to the phenomenal character of colour experience.

Finally, Tye (2000: 150) claims that the phenomenon of colour constancy shows us that the colour of an object is not to be identified with the wavelength of light it reflects in specific lighting conditions. Since the problem of metamers has it that objects with differing spectral reflectances can look the same under certain lighting conditions, the problem of metamers does not challenge colour physicalism that identifies colours with spectral reflectances, or so he thinks. However, this line of reasoning is not convincing because the problem of metamers has it that objects with different spectral reflectances can look the same under some specific illumination conditions. In contrast, colour constancy considers the same object and, thus, the same spectral reflectance under different illumination conditions. Hence, Tyes remarks, true as they may be, do nothing to alleviate the problem of metamerism, which remains as pressing for colour physicalism as ever.

All things considered, the efforts made by Dretske and Tye to defend colour physicalism do not look very promising, and the situation is even worse for the representationalist.[9,10] If representationalism is committed to colour physicalism and there are no persuasive responses to the objections against colour physicalism, it seems that representationalism is fighting a lost cause. Thus, arguing that representationalism is not committed to colour physicalism seems to be the only way out. I will opt for this alternative route and show how this can be accomplished in the following sections.

6. *Physicalism or primitivism about colours? A false choice*

As I have shown in the last section, it seems a desperate move to stick to colour physicalism and defend it against the objections put forth against it. Hence, the representationalist must either rebut the circularity problem and its consequences or show that representationalism might be combined with another theory of colour. As far as the circularity argument is concerned, I do not see any way out for the representationalist but to accept it and the constraint that comes with it.[11] Any attempt to reject the circularity argument is, I think, doomed to failure. Thus, I will take (C) for granted, as do all the representationalists involved in the debate. So, the only remaining option is to defend another theory of colour. Of course, it would, in principle, also be possible to defend colour primitivism against the objections discussed in section 3. But since colour primitivism's problems weigh heavy and colour primitivism runs contrary to the physicalist convictions usually held by representationalists, I will not consider this option.

But what other theory of colour could the representationalist turn to? Upon closer examination, it becomes clear that we have only consid-

[9] For further criticism of the strategies deployed in Tye (2000), see Hardin (2003). For further, albeit quite similar, proposals to defend colour physicalism, see Bradley and Tye (2001).

[10] Another objection against colour physicalism I have not discussed here concerns its dispositional aspect and the resulting problem that colour properties as understood by the colour physicalist have no causal powers. For a response, see Tye (2000: 161–162), and for a critical assessment thereof, see Wright (2003: 520–521).

[11] It is possible to refuse the constraint imposed by the circularity problem by holding that the properties that are represented in colour experience are different from colours. However, such an account would need to say what these properties are that we represent in colour experience and that are different from colours, how they relate to colours and why we mistakenly take them to be colours. Moreover, these properties would need to be characterized without reference on the phenomenal character of colour experience. Otherwise, the problem of giving a circular account would arise again.

ered realist theories of colour so far, i.e., theories assuming that colours are properties that are instantiated in material objects. However, there is no need to accept colour realism in the first place. Neither is representationalism as such committed to colour realism nor is it necessary to accept colour realism to give an adequate response to the circularity argument. Remind that the constraint stated in (C) is all that follows from the circularity problem. And for a theory about colours to satisfy this requirement, it is enough if it does not define colours in terms of subjective responses. Whether colours are properties instantiated in material objects is a separate question. It is open to the representationalist whether to give a positive or negative answer to it when faced with the circularity problem. Thus, the representationalist might willingly adopt a theory of colour that is anti-realist in spirit in the absence of independent reasons for assuming that colours are properties of material objects. So, my diagnosis is that the claim that representationalism is committed to colour physicalism is based on the implicit—and, as I will argue shortly, unjustified—acceptance of colour realism. Therefore, assuming that the choice for the representationalist is between colour physicalism and colour primitivism is entirely misguided.

Before saying something about how to spell out the proposal of adopting anti-realism about colours, it is worth considering the motivation for representationalists such as Tye and Dretske to assume colour realism. It seems natural to them to accept colour realism because they favour tracking representationalism, an externalist version of representationalism. This strand combines the representationalist idea with the claim that mental states obtain their representational content in virtue of their tracking features in the subject's environment. Tracking is cashed out either as a matter of having the function to provide information about the subject's environment (Dretske 1995) or as a matter of being related to the subject's environment in an appropriate way, for example, standing to it in a specific causal relation such as causal covariance under optimal conditions (Tye 1995, 2000). Accordingly, the tracking theory requires that the represented properties be instantiated in material objects, at least in content endowing conditions such as the conditions of evolutionary selection or optimal conditions (Mendelovici 2013). Therefore, tracking representationalism is to be considered externalist in spirit.[12] Obviously, this leads tracking representationalists to claim that colours are properties instantiated in material objects and, consequently, to accept colour realism. Since tracking representationalists are usually drawn to naturalistic metaphysics, they adopt and defend colour physicalism, as shown in the last section.

However, since there is no need to accept the tracking theory in the first place, the representationalist is free to adopt an anti-realist theory

[12] According to Gow (2017), externalist representationalism in general is committed to colour physicalism. She arrives at this conclusion by similar considerations as the ones invoked in this paper.

about colours. Thus, my argument is that representationalism is not committed to colour physicalism, although this comes at the cost of rejecting externalist versions of representationalism such as tracking representationalism.

7. *Anti-realist theories of colours*

The next step for the representationalist is to find out which anti-realist colour theories are on offer and whether they are compatible with representationalism.[13] There are two major candidates: Eliminativism[14] and projectivism. First, eliminativism holds that colours do not exist. According to this view, there are no colours at all, not even uninstantiated ones. Second, projectivism claims that we project the experienced colours onto the objects we perceive due to our having colour experiences but that the perceived objects themselves are not coloured. However, projectivism comes in two versions, literal and figurative projectivism, and they differ significantly (Shoemaker 1990, 1997). Literal projectivism claims that colours are properties that are instantiated in visual fields or other mental entities similar to sense-data (Boghossian and Velleman 1997). In contrast, figurative projectivism has it that colours are properties that are not instantiated at all, neither in material objects nor in visual fields (Maund 2006; Pautz 2006; Wright 2003).

Now, which form of anti-realism is compatible with representationalism? Obviously, eliminativism is no viable option for the representationalist because it denies the very existence of colours—properties that feature in the representationalist's explanation of the phenomenal character of colour experiences. Moreover, literal projectivism is no good option for the representationalist either because it presupposes visual fields or other sense-data-like entities, which conflicts with the representationalist's aim of giving a physicalistically respectable explanation of phenomenal consciousness. Besides these metaphysical worries, literal projectivism is not compatible with representationalism because it holds that colours are modifications of our experiences or mental entities rather than properties that figure in the representational contents of colour experiences. Figurative projectivism, however, is compatible with representationalism because it both assumes the existence of colour properties—as opposed to eliminativism—and it does

[13] Wright (2003) has already argued that representationalism is compatible with the denial of colour realism. However, he does not start his discussion from the circularity argument and does not present anti-realism about colours as an adequate response to the circularity problem. Another major difference is that he claims that externalist versions of strong representationalism are compatible with colour projectivism, whereas I deny this.

[14] The term "eliminativism" is used ambiguously in the debate about the metaphysics of colours. Sometimes, it serves as a label for what I call anti-realist theories, in other cases it is only used to refer to the view that colours do not exist, full stop. I will only use the term "eliminativism" in the latter sense here.

not presuppose the existence of questionable mental entities like visual fields or sense-data—in contrast to literal projectivism.

Though, does figurative projectivism not finally collapse into colour eliminativism since it holds that colours are nowhere instantiated? Not at all. Figurative projectivism assumes that colour properties do, in fact, exist because it claims that they figure in the representational contents of colour experiences. However, it is only a contingent matter of fact that colours are not instantiated in our world. According to figurative projectivism, colours might nevertheless be instantiated in some possible world, e.g., an Edenic world where we would be directly acquainted with the objects around us and their intrinsic qualities (Chalmers 2006). But this is just to say that figurative projectivism is open to the possibility that some possible worlds are different from our actual world concerning the instantiation of colour properties. It does not bear on the central tenet of figurative projectivism that we only mistakenly project colours onto the surfaces of perceived objects when having colour experiences in our actual world.

But what about figurative projectivism as a theory of colour that satisfies the constraint stated in (C)? To begin with, figurative projectivism only tells us something about where colour properties are instantiated or, instead, that they are instantiated neither in material objects nor mental entities. However, it does not tell us anything about the metaphysical nature of colours. It holds that colours are not identical to any kind of properties of material objects since it opposes the very idea that material world objects instantiate colour properties. But what positive claim about the metaphysical nature of colours might be made by the figurative projectivist?

The fact that figurative projectivism holds that we mistakenly project colours onto the perceived objects and that colours are only contingently not instantiated in our world suggests that it is most naturally combined with a primitivist account of the metaphysical nature of colours claiming that colours are simple and intrinsic properties sui generis. Yet, this does not mean that figurative projectivism ultimately collapses into colour primitivism. The primitivist approach described above is realist in spirit, whereas figurative projectivism is anti-realist. They converge in what they say about the metaphysical nature of colours. Both views reject identifying colours with physical properties of external world objects or with dispositions to cause visual experiences with a specific phenomenal character but construe them as simple and intrinsic properties sui generis. Nevertheless, they diverge in what they say about the instantiation of colour properties. To be precise, both primitivism and figurative projectivism deny that colours are properties of our colour experiences themselves. But according to primitivism, colours are had by external world objects, while figurative projectivism holds that nothing instantiates colour properties in our actual world.

8. *Assessing figurative projectivism*

Before dealing with the alleged problems of figurative projectivism, let us first look at how it fares compared to colour physicalism and colour primitivism.[15] In line with what colour physicalism says, figurative projectivism has it that colour experiences are typically caused by the physical properties of the material objects we perceive. However, in contrast to colour physicalism, it denies that colours can be in some way identified with the physical properties of external world objects. Therefore, it disagrees with colour physicalism on whether material objects possess colour properties. This means that figurative projectivism incorporates the advantages of colour physicalism—its compatibility with empirical findings of colour vision—while avoiding its pitfalls—the problems due to identifying colours with the physical properties of material objects.

In accordance with primitivism, figurative projectivism holds that colours are simple, intrinsic, and non-reducible properties. Yet, these two views differ regarding the instantiation of colours. While primitivism purports that material objects have colour properties so construed, figurative projectivism claims that colour properties are nowhere instantiated neither in material objects nor in perceivers or mental entities like sense-data and visual fields. Thus, it is sometimes even held that figurative projectivism, as presented here, is just an anti-realist version of primitivism. So, while figurative projectivism adopts the part of primitivism that fits our common-sense notion of colours—its account of the metaphysical nature of colours –, it does not inherit the problems that come with the claim that colour properties as construed by the primitivist are instantiated in material objects –the unpalatable consequences of colour primitivism regarding what science tells us about colour vision and the surface properties of material objects.

Now, let us examine the difficulties that are supposed to come along with figurative projectivism. As set out in the last section, figurative projectivism states that our colour experiences mistakenly represent objects as coloured, even in the case of successfully perceiving an object. Accordingly, figurative projectivism is committed to the claim that all our colour experiences are non-veridical. However, this flies in the face of our common-sense intuitions about colour perception. It seems appropriate to accept a principle of charity that assumes that not all our colour experiences are blatantly false but that they are more or less correct most of the time (Shoemaker 1997).[16] Moreover, we presume that our judgments about the colours of objects are more or less accurate most of the time and that our discriminations of the colours of objects guide our successful behaviour towards our environment. But figurative projectivism must deny all this. Since figurative projectivism

[15] Some of the points made here are similar to those in Wright (2003: 522).

[16] In contrast, Boghossian and Velleman (1997) hold that it is wrong to appeal to a principle of charity in the case of colour experience.

holds that all of our colour experiences are non-veridical, it comes with the unpleasing consequence that our attributions of colour properties to material objects in our everyday use of colour concepts and terms are wrong, given that we thereby express the contents of our colour experiences. And in addition to that, it does not seem capable of giving us an adequate explanation of how colour perception may serve as a guide for successful behaviour towards our environment when colour perception gets things wrong all the time.

First, let us consider the point about colour perception's alleged uselessness due to its being non-veridical. The assumption upon which this objection is implicitly based is that colour perception can only serve its use in guiding successful behaviour towards our environment if it represents accurately. However, plausible as this claim might seem prima facie, it becomes apparent that it is entirely misguided upon closer consideration. To see this, let us assume that colour experience is misrepresenting all the time. Of course, this implies that colour experience is always non-veridical. Still, it does not preclude colour experience from misrepresenting *systematically*, i.e., that there is some pattern in the way we mistakenly misrepresent objects as having colours. Mendelovici (2013: 422) claims that systematic misrepresentation involves reliability because it "is getting things wrong in the same way all the time." Thus, while being non-veridical, our colour experiences might nevertheless be reliable if they misrepresent systematically. Furthermore, Mendelovici stresses that veridicality must be kept separate from reliability, and it is the latter that secures guidance of successful behaviour. Thus, colour experience might nonetheless serve as a guide to successful behaviour despite misrepresenting our environment if it does so systematically and is therefore reliable.[17]

But what about figurative projectivism and colour experience as systematic misrepresentation? As mentioned above, figurative projectivism holds that colour experience is, at least in the case of perception, caused by the physical properties of the perceived objects. Therefore, as per figurative projectivism, it is plausible to assume that the way colour experience misrepresents objects as having colour properties correlates with the physical properties causing the relevant colour experiences. And this is just what systematic misrepresentation amounts to: a specific type of representation is tokened in similar conditions on various occasions having a content that is never satisfied (Mendelovici 2013: 423). So, while figurative projectivism has it that colour experience is always non-veridical, it can account for the fact that colour perception is useful by claiming that colour experience is reliable because it misrepresents systematically.

[17] A similar point is made in Gow (2016, 2019), who emphasizes that success in not dependent on accuracy. More general, this way of reasoning is usually embraced by proponents of figurative projectivism, see Maund (2006), Pautz (2006) and Wright (2003), for example.

This brings us to the objection that figurative projectivism has the implausible consequence of our everyday colour discourse being flawed because we mistakenly attribute colour properties to material objects. As Boghossian and Velleman (1997: 99) put it, there is no problem in our reports about the colours of objects asserting falsehoods. While it may be true that our everyday colour talk is, strictly speaking, thoroughly false, it can nevertheless be useful in serving the purpose of communication or planning our actions. Boghossian and Velleman point out that even though we often assert falsehoods in everyday talk, e.g., when claiming that the sun rises, there is no need to revise our way of talking about objects and their properties in the light of new (scientific) evidence. Just as we can successfully communicate by asserting that the sun rises and plan our actions based on the belief that the sun rises—though this is false, strictly speaking—we can do so in the case of colours as well. Again, the reason for this is that our colour experience is systematically misrepresenting and, thus, the contents expressed by our assertions about the colours of objects are also systematically false. And since we all systematically misrepresent the objects in our environment with colours that they do not possess, we can successfully communicate with each other even though our colour attributions are false, strictly speaking.

Another worry concerning figurative projectivism issued by Tye (2000: 166) is that it is unclear what it would take for colour experience to be veridical, assuming that figurative projectivism is true. As mentioned before, figurative projectivism has it that it is only a contingent matter of fact that the material objects in our world are not coloured. However, there may be a world where material objects do, in fact, possess the colour properties they are represented as having. In such circumstances, our colour experiences were veridical. This idea can be cashed out, for example, by conceiving of colours as Edenic properties that are instantiated in an Edenic world (Chalmers 2006).

9. *The prospects of representationalism*

In this closing section, I want to briefly summarise what has been said so far and look at the prospects of representationalism. The circularity problem is that representationalism cannot adopt any theory about colours that characterises them as dispositions to produce experiences with a specific phenomenal character. Since colour primitivism faces severe difficulties, it is usually held that the representationalist can only satisfy the constraint imposed by the circularity problem by accepting colour physicalism. Therefore, representationalism seems committed to colour physicalism. However, colour physicalism is not undisputed and defending it from the objections raised against it does not turn out to be fruitful, as I have shown.

But is representationalism committed to colour physicalism at all? I have argued that the answer to this question is “no.” This is because representationalism is not committed to a realist theory of colour that holds that colours are instantiated in material objects. The motivation to adopt a realist theory of colour, so my diagnosis, is based upon defending an externalist version of representationalism, such as tracking representationalism. Yet, since representationalism, in general, can be defended without accepting an externalist theory about mental representation, there is no need to stick to colour realism in the first place. Therefore, the representationalist can deal with the circularity problem by adopting an anti-realist theory of colour. However, this comes at the cost of renouncing externalist versions of representationalism. As far as the choice among anti-realist theories of colour is concerned, the representationalist should opt for figurative projectivism instead of eliminativism and literal projectivism because only figurative projectivism satisfies the requirements of a representationalist theory of phenomenal consciousness. We can thus conclude that the circularity problem does not pose an essential threat to representationalism because representationalism, in general, is not committed to colour physicalism. Only externalist versions such as tracking representationalism are. It is now up to representationalists to develop an updated version of the theory that provides an account of mental representation that leaves externalist commitments behind and shows how an internalist version of representationalism, in combination with figurative projectivism, can be made to work out to give an adequate explanation of colour experiences.

References

Armstrong, D. M. 1997. “Smart and the secondary qualities.” In A. Byrne and D. R. Hilbert (eds.). *Readings on Color. The Philosophy of Color*. Cambridge: MIT Press, 33–46.

Averill, E. W. and Hazlett, A. 2011. “Color objectivism and color projectivism.” *Philosophical Psychology* 24 (6): 751–765.

Boghossian, P. A. and Velleman, J. D. 1997. “Colour as a Secondary Quality.” In A. Byrne and D. R. Hilbert (eds.). *Readings on Color. The Philosophy of Color*. Cambridge: MIT Press, 81–103.

Bradley, P. and Tye, M. 2001. “Of Colors, Kestrels, Caterpillars, and Leaves.” *The Journal of Philosophy* 98 (9): 469–487.

Byrne, A. and Hilbert, D. R. 1997. “Colors and Reflectances.” In A. Byrne and D. R. Hilbert (eds.). *Readings on Color. The Philosophy of Color*. Cambridge: MIT Press, 263–288.

____ 2003. “Color realism and color science.” *Behavioral and Brain Sciences* 26 (1): 3–64.

____ 2007. “Color primitivism.” *Erkenntnis* 66 (1–2): 73–105.

____ 2011. “Are colors secondary qualities?” In L. Nolan (ed.). *Primary and Secondary Qualities: The Historical and Ongoing Debate*. Oxford: Oxford University Press, 339–361.

Campbell, J. 1997. “A Simple View of Color.” In A. Byrne and D. R. Hilbert (eds.). *Readings on Color. The Philosophy of Color*. Cambridge: MIT Press, 177–190.

Chalmers, D. J. 2006. “Perception and the Fall from Eden.” In T. S. Gendler and J. Hawthorne (eds.). *Perceptual experience*. Oxford: Clarendon Press, 49–125.

Dretske, F. I. 1995. *Naturalizing the Mind*. Cambridge: MIT Press.

Gow, L. 2014. “Colour.” *Philosophy Compass* 9 (11): 803–813.

____ 2016. “The Limitations of Perceptual Transparency.” *The Philosophical Quarterly* 66 (265): 723–744.

____ 2017. “Colour hallucination: A new problem for externalist representationalism.” *Analysis* 77 (4): 695–704.

____ 2019. “Everything is clear: All perceptual experiences are transparent.” *European Journal of Philosophy* 27 (2): 412–425.

Hacker, P. M. S. 1987. *Appearance and reality*: *A philosophical investigation into perception and perceptual qualities*. Oxford: Basil Blackwell.

Hardin, C. L. 1988. *Color for philosophers*: *Unweaving the rainbow*. Indianapolis: Hackett.

____ 2003. “A Spectral Reflectance Doth Not A Color Make.” *Journal of Philosophy* 100 (4): 191–202.

Johnston, M. 1992. “How to speak of the colors.” *Philosophical Studies* 68 (3): 221–263.

Lycan, W. G. 1996. *Consciousness and experience*. Cambridge: MIT Press.

____ 2019. "Representational Theories of Consciousness." In Zalta, E. N. (ed.) The Stanford Encyclopedia of Philosophy. Online available: https://plato.stanford.edu/entries/consciousness-representational/. Last accessed on: 25 Aug 2021.

Maund, B. 2006. “The Illusory Theory of Colours. An Anti-Realist Theory.” *Dialectica* 60 (3): 245–268.

____ 2018. “Color.” In Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Online available: https://plato.stanford.edu/archives/sum2018/entries/color/. Last accessed on: 1 Nov 2018.

McGinn, C. 1996. “Another Look at Color.” *The Journal of Philosophy* 93 (11): 537–553.

Mendelovici, A. 2013. “Reliable misrepresentation and tracking theories of mental representation.” *Philosophical Studies* 165 (2): 421–443.

____ 2018. *The Phenomenal Basis of Intentionality*. New York: Oxford University Press.

Pautz, A. 2006. “Can the physicalist explain colour structure in terms of colour experience?” *Australasian Journal of Philosophy* 84 (4): 535–564.

Ross, P. W. 2000. “The Relativity Of Color.” *Synthese* 123 (1): 105–129.

Rubenstein, E. M. 2018. “Color.” In *Internet Encyclopedia of Philosophy*. Online available: https://www.iep.utm.edu/color/. Last accessed on: 29 Oct 2018.

Shoemaker, S. 1990. “Qualities and Qualia: What’s in the Mind?” *Philosophy and Phenomenological Research* 50: 109–131.

____ 1997. “Phenomenal Character.” In A. Byrne and D. R. Hilbert (eds.). *Readings on Color. The Philosophy of Color*. Cambridge: MIT Press, 227–245.

Shrock, C. A. 2017. *Thomas Reid and the Problem of Secondary Qualities*. Edinburgh: Edinburgh University Press.

Smart, J. J. C. 1997. “On some criticisms of a physicalist theory of colors.” In A. Byrne and D. R. Hilbert (eds.). *Readings on Color. The Philosophy of Color.* Cambridge: MIT Press, 1–32.

Tye, M. 1995. *Ten Problems of Consciousness*: *A Representational Theory of the Phenomenal Mind.* Cambridge: MIT Press.

____ 2000. *Consciousness, color, and content.* Cambridge: MIT Press.

Wright, W. 2003. “Projectivist representationalism and color.” *Philosophical Psychology* 16 (4): 515–533.

Croatian Journal of Philosophy
Vol. XXII, No. 64, 2022
https://doi.org/10.52685/cjp.22.64.2
Received: February 22, 2021
Accepted: March 2, 2022

Epistemic Priority or Aims of Research? A Critique of Lexical Priority of Truth in Regulatory Science

JOBY VARGHESE*
Indian Institute of Technology Jammu, Jammu & Kashmir, India

A general criterion for distinguishing between epistemic and non-epistemic values is that the former promotes the attainment of truth whereas the latter does not. Daniel Steel (2010, 2016) is a proponent of this criterion, although it was initially proposed by McMullin (1983). There are at least two consequences of this criterion; (i) it always prioritizes epistemic values over non-epistemic values in scientific research, and (ii) it overlooks the diverse aims of science, especially the aims of regulatory or policy-oriented science. This criterion assumes the lexical priority of truth or lexical priority of evidence. This paper attempts to show a few inadequacies of this assumption. The paper also demonstrates why epistemic priority over non-epistemic values is a problematic stance and how constraining the role of non-epistemic values as 'tiebreakers' may undermine the diverse aims of science.

Keywords: Science and values; epistemic values; lexical priority of truth; non-epistemic values; aims of science.

1. *Introduction*

Recently, the science and values debate has drawn the attention of many philosophers of science, scientists and policymakers. The ideal of value-free science suggests that non-epistemic values such as social,

* I thank Professor Prajit K. Basu, Department of Philosophy, University of Hyderabad, for the useful comments on the earlier versions of this paper. I am also grateful to the anonymous referees and the editor for their valuable comments that substantially improved the paper.

political, moral or economic values should be kept away, or these values have no legitimate roles to play in scientific inference. This ideal has been criticized from different perspectives. Argument from inductive risk is the most significant challenge against the value-free ideal of science. Rudner (1953) argued that since no scientific hypothesis is completely verified, there is always a risk element in accepting or rejecting a hypothesis based on the available evidence. So, value judgments are relevant in weighing the consequences of the mistakes scientists might make when accepting a hypothesis. This line of argument has been further developed by Cranor (1993) and (Douglas 2000, 2009). The gap argument or argument from underdetermination is another critique that is raised against the value-free ideal. The argument states a gap between evidence and a theory (Longino 2002, 2004, 2008). In other words, evidence alone does not determine which hypothesis is true, and the proponents of a value-laden account of science argue that this gap can be bridged by appealing to non-epistemic values (Anderson 2004; Intemann 2005; Biddle 2013; Brown 2013). Similarly, the value-free ideal of science has been criticized by many philosophers of science by arguing that a clear boundary between epistemic and non-epistemic values is necessary to uphold the value-free ideal of science. But drawing a boundary between epistemic and non-epistemic values is not very plausible. So, the defenders of the value-laden account of science put forth a challenge known as the boundary challenge that states that a clear-cut distinction between epistemic and non-epistemic values is not possible (Rooney 1992, 2017; Longino 1995, 1996; Steel 2010; Douglas 2016). The reason is that values such as simplicity, novelty, and ontological heterogeneity might act as both epistemic and non-epistemic values depending on the research contexts. Since a distinction between epistemic and non-epistemic values is not possible, maintaining value-free ideal is also not very plausible in scientific research.

Values in science debate mainly revolve around a very significant question, i.e., how to identify and incorporate a legitimate set of non-epistemic values and eschew the illegitimate influence of such values in scientific inference. Different philosophers of science put forth many suggestions. Douglas (2009) proposes that one should consider the direct and indirect roles of values and these roles will help one evaluate the influence of non-epistemic values. Elliott and McKaughan (2014) argue that when non-epistemic values are involved in scientific research, scientists should be explicit about the role values played in that particular research context. That is to say, the influence of non-epistemic values should be acknowledged and stated as transparent as possible. Intemann (2015) argues that the legitimacy of non-epistemic values can be evaluated by checking whether a particular value or set of values promotes democratically endorsed epistemological and social aims of the research. Steel (2010) argues that all those kinds of influence of non-epistemic values are illegitimate in scientific reasoning when the influence of such values impedes or obstructs the attainment of truth. Steel (2010) and Steel and Whyte (2012) further argue that non-epistemic values should play the role of "tiebreakers" when methodological approaches or two conclusions are equally well defended by epistemic values. In other words, Steel's account allows the "lexical priority of truth/evidence" or "epistemic priority". Two important points follow this principle. Firstly, the epistemic values are characterized in terms of their relationship with truth, and secondly, non-epistemic values should not obstruct the attainment of truth in any scientific inquiry under any circumstances, and if at all they influence scientific inference, their influence must be defended on epistemic terms.

I discuss two important issues in this paper. Firstly, I elaborate and critically analyze Steel's account of values and an underlying assumption that Steel seems to have employed in characterizing the epistemic values. The idea is to demonstrate some of the inadequacies of Steel's characterization of epistemic values as the promoters of truth attainment. This analysis engages with the diverse aims of scientific research, and I attempt to show how these diverse goals provide sufficient place for non-epistemic values to actively participate in different phases of scientific investigations in a legitimate and relevant fashion. Secondly, I will criticize two implications of Steel's proposals: (i) the epistemic values must be characterized in terms of truth and (ii) the influence of non-epistemic values should be limited to only such scenarios in which epistemic values do not completely determine all aspects of scientific reasoning, and when they are involved, they should not conflict with epistemic values. I will argue against these implications and will show that the legitimacy of the influence of non-epistemic values in scientific research need not be always defended in terms of epistemic terms; on the other hand, their influence can be justified in terms of the practical and social relevance of the research.

2. *Values: Characterizations and functions*

A general distinction that is made in science and values debate is between epistemic and non-epistemic values. McMullin (1983) and Steel (2010) argue that epistemic values are acknowledged on the ground that they promote the attainment of truth. McMullin proposes; *"those values that promote the truth-like character of science are epistemic in nature"* (McMullin 1983: 18). Similarly, Steel characterizes epistemic values as that which promotes the attainment of truth or the acquisition of true beliefs (Steel 2010). He further points out; *"Truth should be understood in connection with truth content: a true and very informative belief is more epistemically valuable than a true but trivial belief"* (Steel 2010: 18). However, Steel argues that truth does not necessarily mean true theories.[1]

[1] It seems that Steel is in partial agreement with Catherine Z. Elgin's account of true enough theories. Elgin's claim is that although truth is often considered as a requirement of epistemic acceptability, science and philosophy deploy models,

Non-epistemic values, in general, are such values that are personal, social, economic, moral, religious, or aesthetic in nature.[2] These values are *integral elements in forming the culture and customs of any society, and these values are held to be desirable by different social groups or communities* (Varghese 2021: 237). It is uncontroversial to say that non-epistemic values can function as legitimate determinants in the pre and post epistemic phase of scientific research. But when it comes to the epistemic phase i.e., the justification part of scientific research, there are disputes among philosophers of science regarding the role of non-epistemic values. Some argue that non-epistemic values should be kept away from the epistemic phase of scientific research (Lackey 2007; Sober 2007; Lacey 2010; Betz 2013; Schurz 2013). Douglas (2008, 2009) argues that values can play legitimate roles in scientific inference only if they play indirect roles, for instance, when scientists confront the problem of inductive risk. The argument from inductive risk asserts that scientists are never in a position to have complete certainty about the choice they make in accepting or rejecting a hypothesis (Rudner 1953; Hempel 1965; Douglas 2000, 2009; Wilholt 2009). Inductive risk is the possibility that one may make a mistake in rejecting or accepting a hypothesis that is under study. Douglas makes it very clear that values should not play any direct role in scientific reasoning i.e., they should not *"act as reasons in themselves to accept a claim"* (Douglas, 2009: 96).[3] In general, in the context of inductive risk, non-epistemic values tell us what kind of errors should be preferred and how much evidence is sufficient to make a scientific claim when the claim is likely to bring forth non-epistemic consequences. From an epistemic perspective, accepting a hypothesis when it is wrong is the same error as rejecting a hypothesis when it is true. But ethically speaking, it is not. Here it is also worth discussing how Steel defends the argument from inductive risk. According to Steel, the distinction between epistemic and non-epistemic values can be used to defend the inductive risk argument. Steel starts off by introducing a broad notion of what can be counted as an epistemic value both in an intrinsic and extrinsic manner and further shows how non-epistemic values are worthy candidates to decide upon which kind of error to prefer, that is to say, accepting when a scientific claim is wrong, or rejecting when a claim is right. Steel (2010) argues that Non-epistemic values can influence scientific

idealizations and thought experiments that prescind from truth so that they may achieve other cognitive ends. Elgin's argument is that such felicitous falsehoods function as cognitively useful fictions. They are cognitively useful because they exemplify and afford epistemic access to features they share with the relevant facts (Elgin 2004).

[2] There is a criticism against treating all these values as a uniform group. However, I am not discussing the criticism here since that is beyond the scope of this paper. See Rooney (2017) for the details of the criticism.

[3] This view has been criticized by Elliott and he has given a reformulated version of Douglas's account. See Elliott (2013) for details.

inferences in all those research contexts where epistemic values alone do not decide the activities in different phases of particular research. According to him, the role of non-epistemic values is limited to "tie-breaking" situations. Moreover, Steel emphasizes that the influence of non-epistemic values should not obstruct the attainment of truth.

3. *Steel's characterization of epistemic values*

In what follows, I elaborate on how Steel characterizes epistemic values and in what way Steel's account allows non-epistemic values to play legitimate roles in scientific inferences. Let us start off with Steel's characterization of epistemic values.

3.1 *Values that promote the attainment of truth intrinsically or extrinsically*

Steel distinguishes epistemic values into two categories; intrinsic epistemic values and extrinsic epistemic values. He fleshes out the distinction between them as follows:

> [A] value is intrinsically epistemic if exemplifying that value either constitutes attainment of truth or is a necessary condition for a statement to be true… Epistemic values are extrinsic when they promote the attainment of truth without themselves being indicators or requirements of truth. (Steel 2010: 18)

He suggests that epistemic values can be manifested in different ways, such as through methods, social practices, and community structures, along with theories and hypotheses. One of the most significant features of Steel's theory is his definition of truth. He emphasizes that truth should always be cognized in terms of truth content.

Let us consider Steel's distinction of intrinsic and extrinsic epistemic values. Values like internal consistency and predictive accuracy are intrinsic epistemic values because these values refer to an absence of contradictions or predictions which are true or approximately true. That is to say, these values are the necessary condition for truth. Intrinsic epistemic values are such values that are very robust in the sense of being epistemic in almost any setting. On the other hand, simplicity is an extrinsic epistemic value. The reason is that the world is not so simple and hence, simplicity cannot be considered as a necessary condition for truth. But yet, simplicity promotes the attainment of truth and hence, an epistemic value. For instance, Steel argues;

> Whether external consistency is an epistemic value, however, depends on the truthfulness of the accepted background beliefs … (and) … External consistency might fail to be an epistemic value in a period in which background beliefs are seriously mistaken but become an epistemic value at a later time when the quality of background beliefs has improved. (Steel 2010: 20)

In other words, extrinsic epistemic values such as external consistency are contextual in nature because such values can promote the attainment of truth in a particular context in which they occur. A very significant implication of Steel's intrinsic/extrinsic distinction is that it makes a range of what might count as an epistemic value rather broad. That is to say, many values that are traditionally considered as non-epistemic values can be categorized as epistemic by being extrinsic.

After elaborating the features and the nature of epistemic values, Steel moves on to state the role of non-epistemic values. Non-epistemic values are such values which are not truth-promoting (Steel 2017). According to him, non-epistemic values can play legitimate roles in scientific inferences in scenarios in which epistemic values alone do not fully determine all aspects of scientific investigation. Such scenarios include the choice of methodology, the evidence characterization or the interpretation of data.

Many philosophers of science agree that inductive risk is considerably prevalent in different phases of scientific inquiries (Rudner 1953; Hempel 1965; Douglas 2009; Wilholt 2009). So, the general argument is that non-epistemic values can legitimately influence in assessing which errors are bad and which are worse. The problem that might pop up here is that although non-epistemic values might be necessary to tackle the problem of inductive risk, it might be the case that the set of non-epistemic values which are employed for overcoming this difficulty may not always be legitimate. It is quite possible that the non-epistemic values which may be employed for settling down the issues of inductive risk and underdetermination[4] could be inappropriate in particular research settings. These kinds of inappropriate encroachment of non-epistemic values should be prevented in order to avoid corrupted scientific research, and there should be a criterion to detect whether the influence of a particular set of non-epistemic values is legitimate or not. The principle that Steel suggests as a criterion to distinguish the legitimate influence of non-epistemic values from illegitimate is *the influence principle*.[5] The principle states that non-epistemic values can influence scientific inference epistemically badly if those values act as obstructions in the acquisition of truth. In other words, the influence of non-epistemic values in scientific inference should be in such a way that their influence should not compromise with the epistemic aims. Generally, prediction, explanation and understanding are often depicted as the principal epistemic aims of science. Most importantly, all these aims, in one way or the other, are related to truth or evidence. However, it is not often the case in the context of regulatory science, which is policy-oriented. Regulatory science aims at supporting policy decisions. Pinto and Hicks argue; "*when the goal of conclusive evidence*

[4] In a crude way, one can say that underdetermination involves the idea that models and hypotheses in any particular domain of science are underdetermined by logic and the evidence which are currently available for the models and hypotheses (Longino 1990, 2002; Kourany 2003).

[5] Hicks (2014) terms Steel's principle as *influence principle*. From here on wards, I will also use the same term for further discussion.

conflicts with the practical requirements of regulatory science, regulatory science could legitimately abandon the conclusive evidence standard" (Pinto and Hicks 2019: 3). In what follows, I make an attempt to show that Steel's account is somehow insensitive to non-epistemic goals because of his characterization of epistemic values in terms of truth which assumes the lexical priority of evidence. I will start the analysis of Steel's theory with an assumption which Steel seems to have employed in characterizing the epistemic values as the promoters of truth attainment.

3.2 *Assumption underlying steel's epistemic/non-epistemic characterization and distinction*

Steel's epistemic non-epistemic distinction is construed on the notion of truth. He states that truth should be cognized in relation to truth content and underlines that epistemic values must be characterized in terms of their connection with truth. That is to say, these values should act as the promoters of attainment of truth either intrinsically or extrinsically. It should also be noted that the influence of non-epistemic values is legitimate in only such cases where their influence does not obstruct the attainment of truths which precisely is the influence principle. An implication of the principle is that the influence of non-epistemic values should be justified strictly in epistemic terms which are truth conducive. Hence, it is quite reasonable to think that there is an assumption with which Steel makes the characterization of epistemic and non-epistemic values and their distinction. The assumption can be formulated as follows:

> Assumption (A1): *The aim of science is to provide truth, to be more specific, true beliefs. Epistemically speaking, a belief which has the property of being true is better than a belief that is not true or trivially true, considering all other things equal. Moreover, the value of epistemic justification somehow correlates with truth.*

This assumption is grounded on the idea that truth is the principal epistemic value. This would imply that the ultimate and primary epistemic goal is truth. This assumption appears to be promising in such cases where the ultimate goal of science is always the attainment of truth because; the assumption clearly implies that the ultimate aim of scientific activities is to achieve truth. Moreover, the justification for the acceptance or the choice of a particular model or theory is somehow related to truth. In what follows, I focus on the main problem of Steel's account i.e., the truth-conduciveness in Steel's account, which is committed to a problematic "lexical priority of evidence". The problem becomes more serious when the commitment to truth-conduciveness might lead to the negligence of aims approaches in establishing the role of non-epistemic values in the evaluation of scientific hypotheses or models. In what follows, I present certain possible worries that might follow from allowing the lexical priority of truth or evidence while science deals with multiple aims and attainment of truth might be just one among many aims.

4. *Lexical priority of evidence: Some responses*

I have already outlined the assumption that has been invoked in Steel's characterization of epistemic values. The implications of this assumption call attention to a number of issues that can be raised against Steel's account of values. A very important criticism that is posed against defining epistemic values as the promoters of attainment of truth is the problem regarding the lexical priority of truth. Lexical priority of truth considers truth as the only aim of science and *truth as the absolute value.*[6] Lexical priority of truth eventually leads to the priority of evidence since it is the scientific evidence that will guarantee the objectivity of the scientific research.[7] One of the main reasons why the priority is argued for is because it is said to preserve scientific objectivity intact. However, Brown (2013) shows the necessity of a more nuanced approach when one makes an attempt to show that it is the objectivity that is at stake. The reason is that underdetermination and inductive risk arguments show that there is no values-objectivity conflict. This assertion will put the defender of the value-laden account of science into trouble because, on the one side, they are attracted to the lexical priority of evidence, and on the other hand, underdetermination and inductive risk arguments show that there is no values-objectivity conflict. For instance, Anderson (2004) and Douglas (2009) argue that lexical priority might save scientists from the problem of wishful thinking. Especially for Douglas, the role of values is restricted in assessing the adequacy of available evidence and values, in no way, can be considered as reasons to believe anything.

Brown (2013) further makes a detailed analysis of the problems of priority. He argues that presupposing lexical priority of evidence is not required to argue for underdetermination and inductive risks. The reason is that evidence can turn out to be unreliable or bad sometimes, and in such cases, the priority might lead scientists astray. Moreover, there is no reason to hold the view that when evidence and values pull in opposite directions, we should always follow the evidence if *value judgments are really judgments—adopted for good reasons, subject to certain sorts of tests* (Brown 2013: 837).

[6] As I mentioned earlier, the notion of truth as the absolute value leads to Steel's endorsement of a monist approach which is presented in his influence principle. He fervently argues that the influence of non-epistemic values is legitimate only in such cases where their influence does not impede the attainment of truth.

[7] In a strict sense there is a difference in the lexical priority of truth and lexical priority of evidence. However, in the context of this paper I use them interchangeably for convenience.

5. *Truth as the absolute value: An objection*

This section further explores another objection against the assumption (A1) I mentioned earlier which also presupposes the lexical priority of evidence/truth. I substantiate my arguments based on the aims approach which defends the view that science has got multiple aims and attainment of truth is just one among those many aims. Aims approach also suggest that illegitimate influence of non-epistemic values in scientific inferences can be eliminated by scientists being as much transparent as possible about the goals of their assessments and the roles non-epistemic values played in the assessments as a result (Elliott and McKaughan 2014: 15). Similarly, Intemann (2015) argues that incorporating non-epistemic values should be done in such a way that those values may *promote democratically endorsed epistemological and social aims of research* (2015: 218).

The way scientific investigations are taken up today shows that scientific inquiries are concerned with theoretical aims and pragmatic aims. Theoretical aims focus on extending our knowledge and understanding of the form and contents of the universe. On the other hand, pragmatic aims prioritize the protection of human health and the environment, regulation of chemicals and therapies, informing democratic deliberation, advising policy on climate change, and promoting the capacities of environmental justice and Indigenous communities. Pinto and Hicks (2019) point out that traditionally it was considered that science has just one goal which is *'produce evidences for or against a hypotheses'*. However, regulatory science is policy-related and its goal is not to produce conclusive evidence but to support policy-related decisions. Similarly, Giere (2004, 2006) and Bas van Fraassen (2008) argue that scientific representations can be evaluated in different ways. It can be through the relations that they bear to the world, and sometimes it is in connection with the several uses to which they are put. Since the representations can be evaluated in different dimensions, it is very much plausible to think that the decisions regarding the acceptance of a theory or a model depend on various considerations and truth is only one among the several factors influencing such decisions. Elliott and McKaughan (2014) illustrate this idea very clearly. They say:

> There is an importance of explicitly incorporating a role for agents or users (as well as their goals and purposes) as a crucial component of any adequate analysis. According to this schema, the representational success of models can be evaluated not only in terms of their fit with the world but also in terms of their suitability to the needs and goals of their users. (Elliott and McKaughan 2014: 4)

Here Elliott and McKaughan argue that any object or proposition that is used to represent something else can be analyzed both in correspondence with its fit with the object to be represented and with regard to its fit with the pragmatic functions for which it is employed. In other words, they emphasize the multiple aims of science. An eventual outcome of thinking more carefully about the multiple goals which scientists have when they choose scientific representations is that it enables us to understand how scientists can legitimately prioritize non-epistemic concerns over epistemic ones in certain cases. This prioritization can be seen in various phases of scientific research, such as the choice of the methodology (Varghese 2018) or the assessment of evidential sufficiency (Douglas 2003, 2009). That is to say, scientific models and theories are put to use to represent the world for specific purposes, and it is entirely legitimate to grant that if these models or theories can fulfil those commitments best by forfeiting certain epistemic features for the sake of attaining some of the non-epistemic considerations.

Steel (2010), while discussing the problem of inductive risk, analyses a case study conducted by Cranor (1993, 1995). Cranor's study is concerned with the risk assessment of toxic chemicals when they are exposed to the public. He analyses two models to test the toxicity of different chemicals, which are advantageous in different ways. The first model is more *accurate* than the other model but slower in comparison with the other. On the other hand, the second model is quicker in assessing the toxicity of the chemicals but less accurate compared to the first one. By employing certain mathematical tools to evaluate the risks involved by the use of any of these two models, Cranor shows that if the aim of the research is to minimize the social cost by mitigating the exposure of toxic chemicals to the public, then it is quite plausible to choose the expedited model which is not very accurate in generating the result in comparison with the traditional model but faster in generating the result. Cranor concludes that during this type of risk assessment program, it is legitimate to incorporate non-epistemic factors while choosing between the models.[8] His theory goes as follows:

> Useful risk assessment not only requires drawing reasonably accurate inferences about toxic effects but also demands that those inferences be drawn in a timely manner… (T)he regulatory challenge is to use presently available, expedited, approximation methods that are nearly as 'accurate' as current risk assessment procedures, but ones which are much faster so that a larger universe of substances can be evaluated. (Cranor 1993: 103)

The study is an excellent example that shows that scientific research often aims at achieving certain pragmatic aims rather than mere attainment of truth. In Cranor's case study, there are two important values that play active functions. The first one is about drawing reasonably accurate results and the second one is concerned with drawing inferences in a timely manner. The way science is practised today indicates that there is a clear involvement of pragmatic aims along with epistemic aims in choosing the theories and models in different contexts according to the requirement. The reason why Steel uses Cranor's study

[8] Since my focus is on the social aims of scientific research, and pragmatic aims of science, at least in some sense, are connected to the social aims or policy making, I have used the terms 'social' and 'pragmatic' interchangeably.

in his paper is to show that uncertainties arising from practical challenges faced by specific scientific fields, such as toxicology or climate science, are more than sufficient for nonepistemic values to operate. Imagine that epistemic values normally do place some genuine restrictions on what could and could not be reasonably inferred in a given scientific setting. In such contexts, nonepistemic values might still have room to operate without obstructing epistemic ends. One example of this is relevant to the argument from inductive risk concerns how long one waits and how much evidence one demands before drawing an inference (Steel 2010). In what follows, I will argue that although Steel allows non-epistemic values to play certain roles in scientific research, there are contexts in which Steel's characterization of epistemic values as the promoters of truth and incorporating epistemic priority thesis can be a problematic stance.

6. *Epistemic priority and callousness to non-epistemic goals*

Steel (2010) points out that the argument from inductive risk is often illustrated by such cases where a pressing non-epistemic value, for instance, the protection of human health, provides a powerful reason to draw inferences more quickly, even at the expense of reliability. In such cases, there is a clash between two competing values. On the one hand, there is a model which is more *accurate* (an epistemic feature) but slow in risk assessment and, on the other hand, there is another model which is *expedited* or *faster* (a non-epistemic feature) but not as accurate as of the former. Steel argues that although the choice of the expedited model over the more accurate model appears like a clear case of non-epistemic values directly influencing the choice, which is not the case. He points out that Cranor's study is an example that shows that without compromising epistemic concerns, non-epistemic values might influence scientific inferences legitimately. From an epistemic perspective, the choice between expedited and slower risk assessment methods is a trade-off: quicker inferences versus a somewhat greater chance of error. Steel suggests that Cranor's study is to show that there needs to be a balance between these two *epistemic concerns*[9] and from a purely epistemic perspective, neither of them takes an advantageous position. But when we are concerned with reducing social costs by protecting human health, the expedited method is superior and the best option too. Hence, here the non-epistemic value, protection of human health and thereby reduction of social cost, seems to be playing the role of a "*tiebreaker*." In other words, ease of use and time sensitivity are epistemic values (Steel 2010, 2016) and non-epistemic values such as protection of health will help in deciding between two epistemic values,

[9] Here the two epistemic concerns are quicker inferences and a somewhat greater chance of error.

speed and accuracy. However, I would like to analyze this case from a different perspective in which the social aims of the research may legitimately influence the choice of a model for conducting socially relevant research (Intemann and de Melo-Martín 2010; Varghese 2018, 2019). I will make an attempt to show that Steel's focus on truth is problematic because his account is callous regarding non-epistemic concerns and goals. The callousness that I am going to discuss may need a little elaboration. Although Steel does care for social goals, his characterization of epistemic values focusing on truth is problematic. In other words, when the focus is on truth and epistemic priority, we are setting a boundary for non-epistemic values to engage in scientific research. I shall discuss why setting a boundary is a problem in the last part of the next section. Coming back to the notion of callousness, here it is concerned with the secondary role non-epistemic values should play as 'tiebreakers' when two conclusions or methodological approaches are equally well supported by epistemic values (Steel 2010; Steel and Whyte 2012). Further, scientific practice often incorporates practical or mixed assessments of scientific representations and it is legitimate to prioritize non-epistemic goals when assessing representations in such contexts. Here, I take the discussion forward and argue that the adoption of the aims approach has more potential in achieving social goals than Steel's approach.

The study of Cranor demonstrated that the expedited model of CEPA[10] is more advantageous than the traditional model in assessing the risk if the goal is set to reduce the social costs. In cases like this, it is entirely legitimate to sacrifice some of the epistemic values for the sake of non-epistemic values, for example, sacrificing accuracy for the sake of generating rapid conclusions (although Steel might argue that both these values are epistemic in nature). On the other hand, if the aims of the research were to find the association between exposure to different chemical and adverse health effects for an academic purpose or for publishing the data in a journal for epistemic purposes, then the researchers would have preferred the model which would generate more accurate data. This scenario suggests that the aims of the research may determine which set of values should be prioritized in different contexts.

7. *Aims approach: Right tool for the job*

In varying research contexts, it is de rigueur that researchers should be very specific about those diverse aims which they aspire to achieve. Some of those aims may be purely epistemic in nature, and some of them may not be. The case of toxicity assessment is an example of such

[10] The Committee on Economic and Professional Affairs (CEPA) is associated with monitoring the requisites of the chemical workforce. In addition to that, CEPA members may be asked to review, or act on different materials or information brought to the committee's notice throughout the year.

a research context where the aim is not purely epistemic. Potochnik (2015) and Pinto and Hicks (2019) point out that although traditionally appreciated aims of science included accurate prediction, explanation and representation, other aims have also drawn attention recently. These aims include policy guidance, action within a short time span and facilitating public uptake of scientific knowledge. Elliott and McKaughan (2014) propose that since scientists often have aims that are not purely epistemic in nature, they might choose a model or a theory that is more viable in achieving the aims, and it is even appropriate that certain non-epistemic considerations might be prioritized over epistemic values. However, a worry that pops up here is about the criteria scientists need to employ for appropriately prioritizing non-epistemic values over epistemic ones. How can illegitimate and biased prioritization of non-epistemic values be eschewed? Elliott and McKaughan (2014) try to address this worry by suggesting that scientists must be very *transparent* about the aims and the roles values play in particular research. The transparency can be achieved with the help of *backtracking*.[11] The point is that the prioritization of non-epistemic values must be granted only to the extent that they may promote the goals associated with the assessments that are in play. Another suggestion to avoid illegitimate influences of non-epistemic values comes from Intemann (2015). She argues that incorporating non-epistemic values should be made in such a way that in doing so may promote democratically endorsed social and epistemic aims of the study.

While responding to Elliott and McKaughan's transparency proposal and Intemann's suggestion for democratically endorsed epistemological and social aims proposal, Steel (2017) argues that both these proposals rely on an assumption—epistemic/non-epistemic distinction. He further points out that both Elliott and McKaughan and Intemann, in their arguments at various places, hint that employing a distinction between epistemic and non-epistemic values in practice is very difficult since they are so deeply intertwined. If this distinction is not viable, their proposals also might fall short and face serious repercussions. Moreover, their proposals might also turn out to be unfeasible in such cases where a political majority in a community may endorse such aims which might be incompatible with the integrity of science.[12] So, the final submission of Steel is that a *qualified or a non-absolutist epistemic priority* is necessary for advancing scientific knowledge and human welfare.

A worrying problem of Steel's epistemic priority is that it puts some serious restrictions on science because it allows scientists to consider certain epistemic standards that might sometimes undermine or com-

[11] Backtracking is a concept Elliott and McKaughan (2014) propose to explain how scientists should be transparent about the major assumptions and values involved in an instance of scientific communication.

[12] For more details, refer to Steel's argument with reference to the situation called 'Ibsen predicament' (Steel 2017: 51)

promise scientists' attempts to do socially relevant and responsible science (Brown 2017; Varghese 2021). In the case of policy-oriented or regulatory science such as risk or toxicity assessment, certain restrictions that might be put on the research due to epistemic priority can lead to irresponsible and sometimes even dangerous ways of doing scientific research. According to the epistemic priority thesis, values may only influence science if, in doing so, they respect basic epistemic standards or criteria for what counts as adequate science. Of course, the epistemic priority view accepts that the value-free ideal is not very plausible, but it puts certain restrictions on the roles non-epistemic values can play in scientific inquiry. It is often the case that any decision scientists take in regulatory science may bring forth various societal consequences. As responsible scientists with social commitments, they should make every effort to think through the possible repercussions of their decisions. However, when there is a conflict between values and epistemic standards, always prioritizing epistemic standards can amount to dangerous and potentially irresponsible claims. In other words, the problem with epistemic priority thesis is that it removes the burden of judgment where values and basic epistemic standards conflict. Removal of the burden of judgment is not a good practice in scientific research, at least in the case of policy-oriented scientific research because value judgments are more pervasive in such research contexts. Moreover, the relationship between values and epistemic standards necessarily is more complicated, and hence, the burden of judgment in regulatory science is far more than epistemic priority thesis can tolerate (Brown 2017). But on the other hand, the aims approach provides room for assimilating both non-epistemic values and epistemic standards. Moreover, when this assimilation of values and epistemic standards is not possible, the aims approach will guide researchers to make the trade-off between epistemic and non-epistemic values by considering various social consequences of their decisions rather than focusing on epistemic priority.

8. *Conclusion*

In this paper, I critically examined Steel's characterization of epistemic values as the promoters of the attainment of truth and the functions of non-epistemic values in scientific investigations. A feature of Steel's characterization of values is that they are always assessed in terms of their ability to promote the attainment of truth and it is grounded on the epistemic priority thesis or lexical priority of evidence. An upshot of his thesis is that the epistemic priority thesis or lexical priority of evidence is insensitive to non-epistemic goals and might even undermine diverse aims of science. I argued against this assumption and demonstrated that scientific inquiries are concerned with diverse aims, and the truth is just one among them. I substantiated my claim by advocating the view that models and theories are put to use to represent the world for specific commitments which are either epistemic or non-epistemic and it is entirely legitimate to sacrifice epistemic priority if these models or theories can attend those commitments best by sacrificing some epistemic features for the sake of specific non-epistemic considerations.

In a nutshell, Steel's characterization of epistemic values and the epistemic priority thesis may obstruct the attainment of certain social goals of scientific research. If we grant epistemic priority, then it may place some serious restrictions on science because it allows scientists to always prioritize certain epistemic standards irrespective of the research contexts which may undermine or compromise scientists' attempts to do socially relevant and responsible science. Moreover, certain restrictions due to epistemic priority might also lead to irresponsible and sometimes even dangerous ways of doing scientific research. Hence, I argued that a blend of both epistemic and non-epistemic considerations will nearly always be relevant to the practical needs of users. Thus, it seems that the aims approach is a more viable candidate than Steel's epistemic priority, at least in regulatory science, since the former might guide researchers in making a trade-off between epistemic and non-epistemic values when these values might conflict.

References

Anderson, E. 2004. "Uses of value judgments in science: a general argument with lessons from a case study on divorce." *Hypatia* 19: 1–24.

Betz, G. 2013. "In defense of the value-free ideal." *European Journal for Philosophy of Science* 3 (2): 207–220.

Biddle, J. 2013. "State of the field: Transient underdetermination and values in science." *Studies in History and Philosophy of Science* 44: 124–133.

Brown, M. J. 2013. "Values in science beyond underdetermination and inductive risk." *Philosophy of Science* 80 (5): 829–839.

Brown, M. J. 2017. "Values in science: Against epistemic priority." In K. Elliott and D. Steel (eds.). *Current controversies in values and science.* London: Routledge, 64–78.

Cranor, C. 1993. *Regulating Toxic Substances.* Oxford: Oxford University Press.

Cranor, C. 1995. "The Social Benefits of Expedited Risk Assessments." *Risk Analysis* 15: 353-358.

Douglas, H. 2000. "Inductive risk and values in science." *Philosophy of Science* 67 (4): 559–579.

Douglas, H. 2003. "The moral responsibilities of scientists (tensions between autonomy and responsibility)." *American Philosophical Quarterly* 40 (1): 59–68.

Douglas, H. 2008. "The role of values in expert reasoning." *Public Affairs Quarterly 22* (1): 1–18.

Douglas, H. 2009. *Science, Policy, and the value-free ideal.* Pittsburgh: University of Pittsburgh Press.

Douglas, H. 2016. "Values in science." In P. Humphreys (ed.). *The Oxford handbook of philosophy of science.* Oxford University Press, 609–632

Elgin, C. 2004. "True Enough." *Philosophical Issues* 14: 113–131.

Elliott, K. and McKaughan, D. 2014. "Nonepistemic Values and the Multiple Goals of Science." *Philosophy of Science* 81: 1–21.

Elliott, K. C. 2013. "Douglas on values: From indirect roles to multiple goals." *Studies in History and Philosophy of Science Part A* 44 (3): 375-383.

Fernández Pinto, M., & Hicks, D. J. 2019. "Legitimizing values in regulatory science." *Environmental health perspectives* 127 (3): 03500–8.

Giere, R. 2004. "How Models Are Used to Represent Reality." *Philosophy of Science* 71 (Proceedings): 742–52.

Giere, R. 2006. *Scientific Perspectivism.* Chicago: University of Chicago Press.

Hempel, C. 1965. "Science and Human Values." In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* The Free Press: 81–96.

Hicks, D. 2014. "A new direction for science and values." *Synthese* 191 (14): 3271–3295.

Intemann, K. 2005. "Feminism, underdetermination, and values in science." *Philosophy of science* 72 (5): 1001–1012.

Intemann, K. 2015. "Distinguishing between legitimate and illegitimate values in climate modeling." *European Journal of Philosophy of Science* 5 (2): 217–232.

Intemann, K. and Melo-Martin, I. 2010. "Social values and scientific evidence: The case of the HPV vaccines." *Biology and Philosophy* 25 (2): 203–213.

Kourany J. 2003. "A philosophy of science for the twenty-first century." *Philosophy of Science* 70: 1–14.

Lacey, H. and Lacey, M. I. 201. "Food crises and global warming: Critical realism and the need to re-institutionalize science." In R. Bhaskar et al. (ed.). *Interdisciplinarity and climate change.* London: Routledge, 183–204.

Lackey, R. T. 2007. "Science, scientists, and policy advocacy." *Conservation Biology* 21 (1): 12–17.

Longino, H. 1995. "Gender, politics, and the theoretical virtues." *Synthese* 104 (3): 383–397.

Longino, H. 1990. *Science as social knowledge.* Princeton: Princeton University Press.

Longino, H. 1996. "Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy." In L. Hankinson Nelson and J. Nelson (eds.). *Feminism, Science, and the Philosophy of Science.* Kluwer Academic Publishers, 39–58.

Longino, H. 2002. *The Fate of Knowledge.* Princeton: Princeton University Press.

Longino, H. 2004. "How values can be good for science." In P. Machamer and G. Wolters (eds). *Science, values, and objectivity.* Pittsburgh: University of Pittsburgh Press, 127–142.

Longino, H. 2008. "Values, heuristics and the politics of knowledge." In M. Carrier, D. Howard and J. A. Kourany (eds). *The challenge of the social and the pressure of practice: Science and values revisited,* Pittsburgh: University Pittsburgh of Press, 68–86.

McMullin, E. 1983. "Values in Science." In P. Asquith and T. Nickles (eds.). PSA 1982 II. *Proceedings of the 1982 biennial meeting of the philosophy of science association.*: 3–28.

Potochnik, A. 2015. "The diverse aims of science." *Studies in History and Philosophy of Science Part A* 53: 71–80.

Rooney, P. 1992. "On Values in Science: Is the Epistemic/Non-epistemic Distinction Useful?" In K. Okruhlik, D. L. Hull and M. Forbes (eds.). *Proceedings of the 1992 Biennial Meeting of the Philosophy of Science Association.* East Lansing: 13–22.

Rooney, P. 2017. "The Borderlands between Epistemic and Non-Epistemic Values." In K. Elliott and D. Steel (eds.). *Current controversies in values and science.* London: Routledge, 31–45.

Rudner, R. 1953. "The Scientist qua Scientist Makes Value Judgments." *Philosophy of Science* 20: 1–6.

Schurz, G. 2013. *Philosophy of science: A unified approach.* Routledge.

Sober, E. 2007. "Evidence and value-freedom." In H. Kincaid, J. Dupre and A. Wylie (eds.). *Value-free science.* Oxford: Oxford University Press, 109–119.

Steel, D. 2010. "Epistemic values and the argument from inductive risk." *Philosophy of Science* 77: 14–34.

Steel, D. 2016. "Accepting an epistemically inferior alternative? A comment on Elliott and McKaughan." *Philosophy of Science* 83 (4): 606-612.

Steel, D. 2017. "Qualified Epistemic Priority." In K. Elliott and D. Steel (eds.). *Current controversies in values and science.* London: Routledge, 49–63.

Steel, D. and Whyte, K. 2012. "Environmental Justice, Values, and Scientific Expertise." *Kennedy Institute of Ethics Journal* 22: 163–182.

van Fraassen Bas, C. 2008. *Scientific representation: Paradoxes of perspective.* Oxford University Press.

Varghese, J. 2018. "Influence and prioritization of non-epistemic values in clinical trial designs: a study of Ebola ça Suffit trial." *Synthese* 198 (10): 2393–2409.

Varghese, J. 2019. "Philosophical Import of Non-epistemic Values in Clinical Trials and Data Interpretation." *History and Philosophy of the Life Sciences* 41 (14): 1–17.

Varghese, J. 2021. "A Functional Approach to Characterize Values in the Context of 'Values in Science' Debates." *Logos & Episteme* 12 (2): 227–246.

Varghese, J. 2021. "Non-epistemic values in shaping the parameters for evaluating the effectiveness of candidate vaccines: the case of an Ebola vaccine trial." *History and Philosophy of the Life Sciences* 43 (2): 1–15.

Wilholt, T. 2009. "Bias and Values in Scientific Research." *Studies in History and Philosophy of Science* 40: 92–101.

Croatian Journal of Philosophy
Vol. XXII, No. 64, 2022
https://doi.org/10.52685/cjp.22.64.3
Received: May 3, 2021
Accepted: May 10, 2022

Cladism, Monophyly and Natural Kinds

SANDY C. BOUCHER
University of New England, Armidale, Australia

Cladism, today the dominant school of systematics in biology, includes a classification component—the view that classification ought to reflect phylogeny only, such that all and only taxa are monophyletic (i.e. consist of an ancestor and all its descendants)—and a metaphysical component—the view that all and only real groups or kinds of organisms are monophyletic. For the most part these are seen as amounting to much the same thing, but I argue they can and should be distinguished, in particular that cladists about classification need not accept the typically cladist view about real groups or kinds. Cladists about classification can and should adopt an explanatory criterion for the reality of groups or kinds, on which being monophyletic is neither necessary nor sufficient for being real or natural. Thus the line of reasoning that has rightly led to cladism becoming dominant within systematics, and the attractive line of reasoning in the philosophical literature that advocates a more liberal approach to natural kinds, are seen to be, contrary to appearances, compatible.

Keywords: Classification; cladism; pheneticism; natural kinds; systematics; species.

1. *Introduction*

Cladism is today the dominant school of classification in biology. It incorporates a classification component, a metaphysical component, and a methodological component (Sterelny and Griffiths 1999). The classification component involves the idea that the goal of classification is, or ought to be, to represent phylogeny and only phylogeny, i.e. evolutionary relatedness, or common ancestry. It follows that taxa must be monophyletic (a taxon is monophyletic iff it consists of an ancestor[1] and all and only its descendants; in a monophyletic group each member of the group shares a more recent common ancestor with every other member of the group than they do with any organisms outside the group[2]). The metaphysical component is the claim that all and only the really existing groups or kinds of organisms in nature are monophyletic: if a taxon is monophyletic it is an objectively real group or kind, and if a taxon is not monophyletic it is unreal or artificial, in that it does not correspond to a group with a unified evolutionary history.[3] The methodological component is a set of techniques for inferring phylogeny, the most dominant of which is the Parsimony approach (although other methods, such as the Maximum Likelihood approach, have been preferred by some cladists (Quinn 2017)).[4]

Much of the literature on cladism has focused on its methodological aspect (e.g. the classic discussions of cladism in Hull (1979) and Sober (1988) are almost entirely concerned with this). While recognising that of course the methodological and theoretical components of cladism are not unrelated (the insistence that classification respect only phylogeny would be idle if cladism's methods for inferring phylogeny were unworkable), I propose to focus primarily on the classification and metaphysical components. More precisely, I propose more or less to take for granted the truth of the classification and methodological components, and explore whether, once these are accepted, we must also accept the metaphysical component.

The plan of the paper is as follows. In section 2 I argue that while they have typically been treated as the same question, the classification question and metaphysical question are logically distinct—the answer we give to the former is logically independent of the answer we give to the latter. In section 3, I argue that the characteristic cladist metaphysical position ought to be rejected: monophyly is neither necessary nor sufficient for defining real groups/kinds of organisms. And in section 4, I offer an alternative explanatory criterion for the reality of groups/kinds.

[1] This is 'ancestor', not, as it is commonly stated, 'species', for reasons that will become apparent.

[2] See Podani (2010) for a discussion of the different ways in which monophyly has been understood. He calls the definition I am using the 'consensus' view. Monophyletic taxa contrast with paraphyletic taxa (consisting of an ancestor and some but not all of its descendants) and polyphyletic taxa (sets of species not including a common ancestor of the group). See Ashlock (1971) for an early, useful discussion of these matters.

[3] See e.g. Cracraft, who says that groups lacking a unified evolutionary history are 'nonexistent' (1981, 462).

[4] See Quinn (2017) for a discussion of the many different (sometimes conflicting) meanings 'cladism' and 'cladist' have taken on over the years. Despite these different uses of the terms, the characterisation I offer here (taken from Sterelny and Griffiths 1999) is fairly standard and should be reasonably uncontroversial.

2. *Distinguishing the metaphysics and classification questions*

In the literature on cladism, the metaphysical question and the classification question are typically treated as the same question. That is, the question: which groups should be recognised in classifications, i.e. should be regarded as taxa, is thought to be equivalent to the question, which groups should be recognised as real, natural, objective, groups or kinds in nature?[5] In particular, cladists have held that the view that all and only monophyletic groups are taxa is equivalent to the view that all and only monophyletic groups are real.

But the questions are logically distinct. For instance, many theorists hold that species are real, objective units in nature, whilst higher taxa—families, classes and the like—are 'constructs of the systematist's mind, not existing in nature in any real sense' (Eldredge and Cracraft 1980: 250).[6] Those who hold this view do not, typically, hold that species are the only taxa. They may recognise that higher taxa do have a role in classification. It is just that as such, they do not correspond to really existing units in nature. In particular, it would seem to be perfectly consistent for one to be a cladist about classification while accepting the popular view that only species, not higher taxa (even if they are monophyletic), are objectively real (indeed this combination of views is explicitly defended by some cladists e.g. Eldredge and Cracraft (1980)). Of course cladists about classification have tended to accept the traditionally cladist view about the latter question, according to which all and only monophyletic groups are real, whether species or not. But this is not, I suggest, compulsory once one has accepted the cladist view on classification.

Conversely, one may hold that certain groups or kinds are real, without holding that they are taxa (as I shall discuss below). So it would seem that, conceptually, being a real group or kind is neither necessary nor sufficient for being a taxon. Of course one *may* hold that all and only taxa are real groups; but this would be a substantive position, it does not follow analytically from the concepts of 'taxon' and 'real group/kind'. One who recognises taxa they do not believe are real are not conceptually confused, I maintain.

In the context of cladism, the classification question is: given a phylogeny, is it the case that the taxa that are recognised by the correct classification are all monophyletic? While the metaphysical question is: is it the case that the only groups of organisms that are objectively real are the monophyletic taxa?

[5] See e.g. Cracraft (1981: 459).

[6] See Mishler and Donoghue (1982). Often this is expressed in ontological terms: species are individuals, higher taxa are collections of species, and thus 'classes' (Eldredge and Cracraft 1980).

It will be helpful to distinguish three views on classification from three views on ontology:

Classification:

1. All and only monophyletic groups are taxa
2. Taxa may be monophyletic or paraphyletic (but not polyphyletic)
3. Taxa may be monophyletic, paraphyletic or polyphyletic

Ontology:

4. All and only monophyletic groups are real[7]
5. Real groups may be monophyletic or paraphyletic (but not polyphyletic)
6. Real groups may be monophyletic, paraphyletic or polyphyletic

Characteristically, (1) and (4) have been held by cladists;[8] (2) and (5) by evolutionary taxonomists (see e.g. Mayr (1942), Simpson (1961));[9] and (3) and (6) by pheneticists.[10] The claim I defended above about the logical independence of the classification question and metaphysical question can be understood as the claim that this traditional combination of views is not logically compulsory. If they are logically distinct *one may combine any of the views on classification with any of the views on metaphysics.* Some of these combinations would be odd—e.g. combining the phenetic view on classification with the cladist view on metaphysics; odd but not perhaps logically contradictory. One may combine the cladist view on classification with the phenetic view on metaphysics less oddly perhaps. But defending the consistency of *all* of the positions on taxonomy with *all* of the positions on metaphysics is not required for my argument. All that is required is that the cladist view on classification be consistent with all three positions on ontology.

As I have noted, this consistency has not been generally recognised. It has been assumed that the classification question *just is* the metaphysical question.[11] Once we distinguish the questions, it still remains

[7] Of course, here and throughout the paper this should be understood as referring to the question of which groups *of organisms* count as real groups or kinds.

[8] In the philosophical literature (4) has been defended by Rieppel (2005).

[9] Evolutionary taxonomists allow paraphyletic but not polyphyletic groups because they believe classification (and metaphysics) ought to represent and take account of divergent, but not convergent evolution (Ridley 1986).

The group comprising lizards and crocs but excluding birds is paraphyletic. Birds and crocs are more closely related to each other than either is to lizards, so grouping crocs and lizards together apart from birds can only be justified on phenetic grounds: by the fact that crocs and lizards are more similar to each other than either is to birds. This is the case because birds have diverged morphologically from other members of their clade. But convergence is not respected by the second view. So in some cases where evolutionary relatedness clashes with overall similarity (ones deriving from divergence) the view opts for the similarity criterion; in other cases where they clash (ones deriving from convergence) it opts for evolutionary relatedness.

[10] The metaphysical positions are not as explicit in evolutionary taxonomy and pheneticism as in cladism.

[11] For instance, Sober, in his characterisation of cladism (1988), only mentions the classification and methodological components, presumably because he takes it that the classification component encompasses the metaphysical component.

the case of course that cladists *have in fact* defended (1) and (4). But my claim is that they needn't have done so; that accepting (1) does not logically compel them to accept (4). This ought to be an agreeable fact for cladists given that, as I will argue below, (4) is very implausible.

It is not surprising that the metaphysical and classification questions have not generally been distinguished. It is often said indeed that the aim of biological classification is to identify 'natural' groups (Ridley 1986). The goal is the construction of a 'natural' classification that identifies and names all and only the real, objective groups and kinds in the area under study, one that 'cuts nature at its joints', i.e. the distinctions it draws correspond to real, objective, mind-independent divisions between things in the world. On this view, the classification question and the metaphysical question go together: in a natural classification, a group is a taxon iff it is a natural group or kind. And certainly there is a sense of 'classification' on which this is reasonable: on which there is no meaningful distinction between classifying and identifying kinds. Nonetheless, I think the question of classification can be and often is understood in a different sense, a sense in which it is an open question whether the groups picked out by a (the?) correct and objective classification system are all and only the natural or real groups or kinds in nature. We can accept that the aim of a classification is to carve at joints, and mark objective distinctions in nature. For instance, Ridley says that an 'objective classification' is one in which 'the choice of characters is dictated by a theoretical principle. The principle must specify some discoverable hierarchical property of nature, which it is desirable and technically possible for classification to represent' (1986, 3). Cladism arguably satisfies this condition in its aim of representing the objective branching order of the tree of life. If humans and chimps are more closely related to one another than either is to gorillas, this is an objective fact about the world in a way that relations of similarity can never be. Hence cladism's (in my view) justified claim to being a more objective, and thus more adequate, system of taxonomy than either pheneticism or evolutionary taxonomy. But it is quite another thing to expect of a classification that it identify all and only the really existing groups or kinds in nature. It is far from obvious that the reasonable requirement that a classification be 'objective', or 'natural', should be interpreted as the requirement that such a classification should achieve this much stronger and more ambitious aim.

More specifically, I will understand biological classification in a relatively minimal sense, as involving an objective, non-arbitrary, unambiguous system of organising, grouping, ranking and naming. In biology we expect a classification to be hierarchical, i.e. involve classifying into ever more inclusive, non-overlapping categories. Two points are important here. First, such a system must respect natural divisions *sensu* Bird (2018), in the sense that it maps *only* natural divisions among organisms; it need not map *all* the natural divisions: this would be asking too much. Secondly, biological classification need not pick out all and only natural *kinds*. I follow Bird (*ibid*) in claiming that natural divisions are necessary but not sufficient for natural kinds. A classification may identify and name taxa that are not natural kinds, and there may be natural kinds (involving natural divisions) that it does not identify or name.[12] (These points will become clearer in due course.)

One uncontroversial way of distinguishing the classification and metaphysics questions would be to argue that classification is or ought to be pragmatic, i.e. relative to human interests and purposes (scientific and/or non-scientific), such that a classification system need not identify all and only the real groups or kinds in nature (Dupre 1981, 1993). It is important to see that this is not the view I am defending. I am suggesting that even if we accept (as I think we should) that a classification system ought to be objective—ought to capture objective divisions in nature—it still may not identify all and only the real groups or kinds.

The logical independence of the classification and metaphysics questions is implicit in Sterelny and Griffith's (1999) discussion of cladism. On the classification question, they side with cladism (196–197). They reject pheneticism, as well as the compromise, or 'mixed' approach to classification favoured by evolutionary taxonomists, on the standard grounds that of the three systems, only the phylogenetic approach has a chance of being systematically objective, in that what it aims to capture—the order of evolutionary branching and thus what Darwin called propinquity of descent—is genuinely objective, whereas both pheneticism and evolutionary taxonomy must appeal to judgments of similarity and extent of evolutionary divergence, which can never be rendered fully objective.

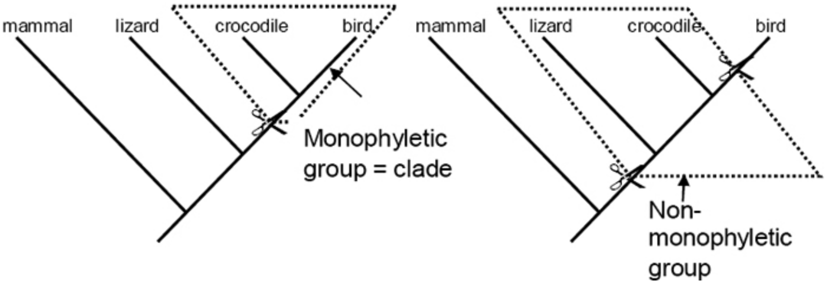
But on the metaphysical question, they adopt the compromise (characteristically evolutionary taxonomy) view. 'To the extent that cladists really do want to reject truncated monophyletic [i.e. paraphyletic] groups—groups that contain nothing but a single species' descendants, but not all of them—their views are too extreme' (198). This is because, they think, there are real groups that are paraphyletic: 'We think it quite likely that there can be good evolutionary hypotheses about such *paraphyletic* groups. For example, there may well be sensible evolutionary hypotheses about all the nonmarine mammals… it's easy to imagine events that affect all of, and only, that truncated group.' (198) Note the implicit criterion for recognising groups—are there good/sensible evolutionary hypotheses about them? Are there events that affect all and only their members? I will return to this. Although they don't explicitly present it this way, I take it that Sterelny and Griffiths are accepting the cladist position on the classification question, while accepting the evolutionary taxonomy view on the metaphysical question, on the grounds that we use different criteria to determine a taxon and

[12] Thank you to an anonymous referee for encouraging me to be more explicit about what I take classification to be.

to determine a real group: the criterion for the former is evolutionary relatedness (phylogeny); the criteria for the latter at least includes whether there are good evolutionary hypotheses about the putative group. Even in Sterelny and Griffiths, the distinction between the metaphysical and the classification question, and the possibility of accepting the cladist view of classification while rejecting the characteristically cladist metaphysical view, is only implicit. But they must accept the distinction, if they think the compromise view on the classification question is untenable (196), but think also that we should recognise paraphyletic groups. This only makes sense if these are addressing different questions, that is, if being a taxon is not the same thing as being a real group. In particular, it follows that even if the compromise view of classification must be rejected, the compromise view on the metaphysical question may still be accepted.

Below I will argue that once we have accepted paraphyletic real groups, there is no justification for stopping there: we can, and perhaps should, also accept polyphyletic real groups. That is, the compromise view on the metaphysical question is unmotivated, and we should adopt the characteristically phenetic view on the metaphysical question (which, recall, is the view that real groups may be either monophyletic, paraphyletic, or polyphyletic), though not on traditionally phenetic grounds.

To conclude this section, consider the well-known phylogeny of birds, crocs and lizards:



The cladist about classification holds that birds and crocs should be grouped together apart from lizards, while evolutionary taxonomists would group lizards and crocs together apart from birds. But this, I suggest, is entirely a question concerning classification. It is a further, distinct question whether the group including crocs and birds but excluding lizards is objectively real in a way that the group including crocs and lizards but excluding birds is not. To put it another way, according to cladists, classification is all about the sister-group relation. Crocs and birds are sister groups relative to lizards. Birds/crocs and lizards are sister groups relative to mammals. But it is hard to see why the sister-group relation should tell us anything very much about the metaphysics of real groups or kinds.

3. *Against monophyly as a metaphysical criterion*

So cladists (about classification) *may*, logically speaking, reject the traditional cladist view on the metaphysical question. In this section I will argue further that they *should* reject it. Monophyly is, I will argue, neither necessary nor sufficient for a group of organisms to count as a real group or kind.

The cladist metaphysical criterion is notoriously strict; too strict, according to many. There are arguably real groups that, because non-monophyletic, it does not count as real. I concur with this judgment. But I will argue that it is also too liberal: it counts too many groups as real. In short, some real groups are not monophyletic, and some monophyletic groups are not real. Since being monophyletic is neither necessary nor sufficient for being real, the criterion should be rejected.

3.1 *Questioning the necessity*

The most obvious sense in which the monophyly criterion of reality is too strict is that it rules out all ancestral, that is, non-monophyletic species. I discuss this in the following section (3.1.1). Setting species aside for the moment and focusing on higher taxa, it has seemed to many that in ruling out the reality of certain higher taxa counted as real by commonsense and received taxonomic theory—reptiles, fish, dinosaurs (minus the birds), great apes (minus humans) etc.—because paraphyletic, cladism is committed to the 'absurd' conclusion that 'there is no such thing as a fish/reptile/dinosaur/ape…' Whether or not this is indeed absurd, or just a somewhat surprising consequence of an otherwise sound taxonomic philosophy that we can and must learn to live with, the point I wish to make here is that it has been assumed that in adopting the cladist view of classification, and thus refusing to admit paraphyletic taxa, the cladist is thereby committed to rejecting the reality of non-monophyletic groups, as the classification question and metaphysics question have not been distinguished. If I am right that these questions are distinct, and accepting the cladist answer to the former does not entail accepting the monophyly criterion for the reality of groups, it follows that in refusing to accept reptiles, fish etc. as taxa, the cladist need not deny that they form real groups, and thus *need not* embrace the 'absurd' conclusions. For the conclusion follows from the rejection of non-monophyletic real groups, not the rejection of non-monophyletic taxa. As Sterelny and Griffiths note, the view that 'there is no such thing as a reptile' follows directly from the cladist *metaphysical* thesis—it follows from the claim that reptiles do not form a real group. One could it seems hold that there is no reptile taxon, yet hold that reptiles are a real group, and thus that there are reptiles, just as Sterelny and Griffiths appear to hold that there is no terrestrial mammal taxon (as they accept the cladist view on classification) but there is a terrestrial mammal real group (see Devitt 2011).

I do not here propose to offer a verdict on the reality of particular paraphyletic groups. In the final section I will suggest a criterion for reality that may be used to decide on such questions. My point here is simply that if we reject the strict cladist metaphysical view then we are not committed to denying the reality of taxa such as reptiles, fish and so on merely on the grounds of their non-monophyletic character. Whether these traditional taxa, or other paraphyletic groups, are real groups or kinds will depend on whether they satisfy the criteria I will outline in the final section; the point here is just that we are not *compelled* to rule them all out automatically just on the grounds that they are not monophyletic.

In this context it is worth considering Griffiths' suggestion in an earlier paper that 'reptiles' is example of reference failure, because the reptile taxon is paraphyletic, and thus there is no real division in nature corresponding to it (1994: 210).[13] On the view I am defending it *might* be correct to say that there is reference failure here, but not because the group is paraphyletic. On my view paraphyletic groups *can be* real but often are not. Whether 'reptile' names a real group (and thus whether or not it refers) depends not on whether it is monophyletic or paraphyletic, but on whether it is explanatory (I will say more about this criterion in the final section). And it is worth noting that Griffith's position here—that all paraphyletic groups are unreal—conflicts with his and Sterelny's position (*ibid*) that some paraphyletic groups, such as terrestrial mammals, are real. Thus 'terrestrial mammal' presumably refers, despite referring to a paraphyletic group.

3.1.1 *Species and monophyly*

Species have always presented a problem for cladism, on both the classification and metaphysics fronts, given that to the extent that species may be ancestral to other species, they may fail to be monophyletic (Sober 2000: 166). Different species concepts will have different implications about when and why species may fail to be monophyletic. For instance, on Mayr's Biological Species Concept (BSC), which defines species in terms of interbreeding and reproductive isolation, one interbreeding population may give rise to another from which it is reproductively isolated. These would each count as separate species despite the parent species being paraphyletic (Ereshefsky 1998: 105–106). Cladists tend to adopt one or other of the various historical species concepts, either a version of Simpson's evolutionary species concept, according to which 'a species is a lineage evolving separately from others and with its own unitary evolutionary role and tendencies' (Wiley 1992), or a version of the phylogenetic species concept, according to which a species is a branch of the phylogenetic tree, beginning at a speciation (branching) point, and terminating either at another speciation point, or at the

[13] See also Rieppel, who argues that 'Reptilia' doesn't designate a natural kind because it is not monophyletic. It is rather an 'artificial' kind (2005: 467).

extinction of the lineage.[14] But even on these species concepts, species will, on the face of it, still be paraphyletic, if they have any descendants.[15] Even if, as the phylogenetic species concept states, new species may not arise through phyletic evolution in a lineage without splitting, but may only arise through branching (subdivision of an existing lineage), it will still be the case that some species will be ancestral to others, and thus will be paraphyletic. Of course cladists are notoriously wary of the ancestor-descendant relation. But phylogenetic cladists do have to accept, as an ontological claim, that there are such things as ancestral species that give rise to daughter species. Their point is the purely epistemological (and reasonable) one that we can never *know* on the basis of the evidence which species have been ancestral to others.

The uncontroversial case in which admitting the existence of ancestral species conflicts with the cladist principles is where a species continues to exist after budding off a daughter species (as some cladists e.g. Wiley (1992), and others sympathetic to cladism e.g. Hull (1979), think can happen). The parent species will then be paraphyletic, and thus illegitimate: after the split, there will be organisms/populations in the parent species that are more closely related to (share a more recent common ancestor with) organisms/populations in the daughter species than they are to organisms/populations in the earlier phase of the parent species before the split, yet are being classified with the latter and not with the former (just as, in the case in which the stem species does go extinct when the lineage divides, so that species *a* gives rise to species *b* and *c*, *b* and *c* are grouped together in the cladogram apart from *a*: the group *a* and *b*, apart from *c*, would be paraphyletic). Yet even if we follow Hennig and other cladists in their view that a species always goes extinct when it splits, it will still seemingly be the case that the parent species will be paraphyletic, as we are excluding from it some of its descendants.[16]

Hennig originally intended his criterion of monophyly only to apply to supra-specific taxa (Ereshefsky 1998). Later cladists went to the opposite extreme and merely *assumed* species were monophyletic, which assumption underlies the popular definition of monophyly: a species along with all (and only) its descendants if it has any. On this view '(s) pecies are taken to be monophyletic *a priori*' (Brandon and Mishler 1987: 118). Subsequent cladists, such as Brandon and Mishler, urged that species need to be, as it were, *internally* monophyletic. After all, if a species comprises, say, three disjoint populations, and does not in-

[14] There are several phylogenetic species concepts (Baum and Donoghue 1995; Wilkins 2009), but the differences between them are not important for our purposes.

[15] This hasn't always been recognised, for instance Ereshefsky in his (1998) seems to suggest that there are no paraphyletic ancestral species on the phylogenetic species concept, as do other cladists: see below.

[16] Ridley disagrees (1989). He suggests that in such a case, the species that goes out of existence at the point of branching counts as monophyletic. I criticise this view below.

clude the common ancestor of those populations, the species will be non-monophyletic even if terminal. Thus Brandon and Mishler suggested replacing the above definition of monophyly with the following definition: 'A monophyletic taxon is a group that contains all and only descendants of a common ancestor, originating in a single event' (118). The common ancestor here is thought to be an individual organism or local population (118-119). One consequence of this conception is that populations below the species level may be monophyletic, though the species is the least inclusive monophyletic *taxon*.[17]

This shift in perspective to a more fine-grained understanding of monophyly is well motivated. But the problem of ancestral species remains, as we shall see. In this section I will survey some attempts to reconcile species with the principles of monophyly.

Ridley (1989) accepts the cladist classification principle (that all and only monophyletic groups are taxa), but argues that all species satisfy it on the cladistic (phylogenetic) species concept. Other species concepts, such as the BSC, fail to satisfy it. The BSC allows paraphyletic taxa, because in the case when a species splits, with one branch diverging and the remaining branch remaining much the same, while the cladistic concept (as he understands it) says the unchanged species has become a new species at the branch point, the BSC says it remains the same species, as former and later segments could potentially interbreed (13). Ability to interbreed is not sufficient for conspecificity on the cladistic concept. On the cladistic concept, species are monophyletic in the sense that they are monophyletic *up to the next speciation event* (if there is one). All the descendants of the species are included in the taxon *so long as no speciation takes place.* This is a bit like saying my grandfather is alive because he was alive up to the point when he died. It's true that paraphyletic taxa are monophyletic if you ignore the branches that make them non-monophyletic.

Ridley's view is that only if the parent species continues to exist after budding off a daughter species does it count as paraphyletic; if it goes extinct at the point of branching it counts as monophyletic. 'The species before and after the split are different branches of the phylogenetic tree, and both branches are monophyletic.' (13) Again, this seems to involve an unmotivated revision of the standard understanding of monophyly: an ancestor along with *all* and only its descendants. Even if a species ceases to exist at the point at which it gives rise to descendant species, insofar as it has descendants, the taxon consisting of that species minus its descendants is paraphyletic.

Brandon and Mishler, in their influential (1987);[18] similarly argue that species are monophyletic on their version of the phylogenetic spe-

[17] If this is accepted, we would need to revise the cladist classification principle, since it is no longer the case that all monophyletic groups are taxa.

[18] They follow Mishler and Donoghue (1982); see also Donoghue (1985) for similar position.

cies concept, according to which a species is 'the least inclusive taxon into which organisms are grouped due to monophyly'. (Monophyly is their grouping criterion, while they adopt a pluralistic ranking criterion, to accommodate the plurality of evolutionary forces responsible for making species into coherent and separate lineages. Monophyly is only the grouping criterion because taxa other than species can be monophyletic; thus being monophyletic is necessary but not sufficient for being a species.) But 'the least inclusive monophyletic group' can only apply to species as terminal taxa. Ancestral species are, as we have seen, not monophyletic.

Brandon and Mishler attempt to get around this problem by denying that any species are ever ancestral. Only smaller units (e.g. organisms or populations) are ancestral to species. Their point seems to be that the full implications of the rejection of anagenetic speciation have not been understood, inasmuch as the idea of species being ancestral to other species has been retained in a cladogenetic setting. But with the acceptance of the idea of speciation by splitting, the idea of ancestral species can be rejected. This doesn't appear to solve the problem of ancestral species however. Take the individual or population X that is considered the 'ancestor' of all members of monophyletic species S in Brandon and Mishler's analysis. X did not spring into being from nowhere; it itself descended from ancestors. Those ancestors belonged to a different species, *ex hypothesi*. Call it S*. The members of S*, let's suppose, all descended from a common ancestor, X*. So S* contains only descendants of X*. But S* does not contain *all* the descendants of X*, since it does not include the members of S. Thus S* is not monophyletic.

The theorists I've been discussing can only continue to uphold the cladist metaphysical and classification principles if they revise the definition of monophyly to include all phylogenetic species (species as understood on the phylogenetic species concept) by definition. Instead of defining a monophyletic group as 'an ancestor and all and only its descendants' we would have to define it as follows:

A taxon is monophyletic so long as it satisfies one of the following conditions:

1. It is a phylogenetic species
2. It consists of an ancestor plus all and only its descendants

So in the case of a stem species *a* budding off two terminal daughter species *b* and *c*, rather than there being three monophyletic groups as per usual—*b*, *c*, and *a*+*b*+*c*—there would be four: *a, b, c,* and *a*+*b*+*c*. Such a revision would appear ad hoc, if motivated in no other way than by a desire to maintain the cladist principles. The alternative is to accept that ancestral species are non-monophyletic, and revise the cladist principles accordingly.

Eldredge and Cracraft, in their classic text (1980), accept the point I have been urging against cladists such as Ridley, that ancestral species cannot be monophyletic (90). They note that a strict application of cladistic principles would require all taxa to be terminal (as all taxa must include every descendant species in order to be monophyletic). Thus if we are to accept some ancestral taxa, cladist principles would need to be modified.

Eldredge and Cracraft are robust realists about species. Throughout their book they defend the view I have adverted to above, that species are ontologically real, discrete, objective, mind-independent units in nature (particular, concrete things, or individuals), while higher taxa are subjective and more or less arbitrary projections of our minds. Thus if, as they accept, ancestral species are non-monophyletic, we have here a clear counterexample to the cladist metaphysical principle: it's not the case that only monophyletic groups are real. (They are also implicitly rejecting the view that all monophyletic groups are real, in their view that higher taxa are conventional projections of our minds.)

At times however they appear to wish to continue to defend the traditional cladist metaphysical principle. So they claim elsewhere in the book (266) that non-monophyletic groups are 'non-existent', which would imply that ancestral species are non-existent, which directly contradicts their above-mentioned realism about all species. This illustrates the tension that exists in cladist thought with respect to this question. Cladists cannot say both that all species are objectively real, and that only monophyletic groups are real. One of these has to give way.

The best solution, I would suggest, is to reject monophyly as a necessary condition of reality. All phylogenetic species are real, including those that are paraphyletic, and thus it's not true that only monophyletic groups are real.

Christofferson (1995) accepts that ancestral species are not monophyletic and that this creates a *prima facie* problem for traditional cladism (446–447). His response is that there are fundamentally two (equally real and important) types of taxa, species and monophyletic higher taxa, and these belong to quite different ontological categories. Species are understood dynamically as evolving lineages (we take a *transformational* view of them), while monophyletic higher taxa are understood statically as hierarchically organised sets of taxa (we take the *taxic* view of them). 'Phylogenetic systematics involves integration of these two world views [the transformational and taxic] by recognition of two ontological kinds of taxa: species, which are continuous strings of ancestor-descendant populations ranked serially (the transformational approach), and monophyletic taxa, which are discontinuous taxa ranked hierarchically (the taxic approach)' (444). Thus species are exceptions to the strict cladist metaphysical principle.

It would seem to be an implication of Christofferson's view that no species, even terminal species, are ever monophyletic. Treating any species as monophyletic is akin to a category error. This is a return to Hennig's original view. I would argue that terminal species can be monophyletic if they satisfy Brandon and Mishler's conditions on monophyly. But Christofferson is right (as against Brandon and Mishler and Ridley) that (a) there are ancestral species, and (b) they are non-monophyletic, and thus we need to revise the cladist metaphysical principle.

I have been focusing on the need to revise the cladist metaphysical principle to accommodate realism about species. But of course if ancestral species are non-monophyletic, they are also a counterexample to the cladist classification principle (all and only taxa are monophyletic). If ancestral species are taxa, then the cladist classification principle would need to be modified. The only other option would be to deny that ancestral species are taxa. This may seem like a radical proposal, but it is a straightforward implication of, for instance, the definition of species taxa advanced by Mishler and Donoghue: 'a species is the least inclusive taxon recognised in a classification, into which organisms are grouped because of evidence of monophyly' (1982), or that advanced by Mishler and Theriot: 'taxa are ranked as species because they are the smallest monophyletic groups deemed worthy of formal recognition' (2000, quoted in Wilkins 2009: 213). If these definitions are accepted, ancestral species are not species taxa. The only species taxa are terminal species (species that are either extant, or went extinct without speciating). This appears to have been Hennig's view (1966; see Richards 2016: 163). Hennig suggested there were no stem species taxa apart from the entire clades they gave rise to, that is, a stem species is identical to the entire clade it is the stem species for: 'in the phylogenetic system [the stem species] … is equivalent to the totality of species in the group' (1966, quoted in Richards 2016: 163). Similarly, Mishler and Donoghue (1982) raise the possibility of peripheral isolate-type allopatric speciation, where the parent species would be paraphyletic (499). Their solution is that in such a case we should say that the parent species is not in fact a species at all. In other words, since species cannot be paraphyletic, there are no ancestral species. All species are either still living, or went extinct without branching.

I have suggested above however that ancestral species are real groups (and thus that the cladist metaphysical principle should be modified). The notion that ancestral species are real groups but are not taxa may seem strange, but one of the main themes of this paper is that the issue of metaphysics and the issue of classification should be kept distinct. I am suggesting that there are likely to be a wide range of real groups that are not monophyletic, so are not taxa, if we accept, as I think we should, the cladist classification principle. Sterelny and Griffiths' terrestrial mammals are an example. They are a real group on their criterion for reality, but do not count as a taxon on their cladist criterion of classification. Ancestral species would just be just another example, and do not seem to raise any special, further difficulties.

3.2 *Questioning the sufficiency*[19]

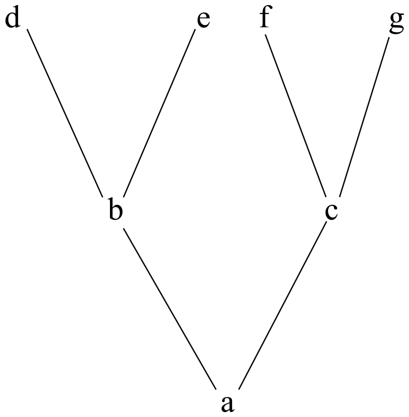
Whether or not it is too strict in ruling out some real groups, the monophyly criterion of reality is arguably too liberal. It holds that every monophyletic taxon is a real group. The first potential worry is that this commits one to the reality of a vast number of groups or kinds. If monophyly is a sufficient condition for the reality of groups we end up with a proliferation of real groups, organised hierarchically, that may seem metaphysically profligate. There is after all a separate monophyletic group for every species that has ever lived (assuming all species are monophyletic in Brandon and Mishler's sense): the group consisting of that species along with its descendants if it has any, or that species alone, if it doesn't.

Ridley (1993: 369-70) notes the huge number of evolutionary branching points in the history of life. Each represents a distinct monophyletic clade, so are all equally taxa for the cladist, but there are obviously far too many to all be given a Linnaean rank (see also Eldredge and Cracraft 1980: 221, Ereshefsky 1997: sect. 3). He argues that this doesn't matter because Linnaean ranks are subjective and conventional anyway, so in assigning them we can ignore lots of 'real' taxonomic levels. Each monophyletic clade is a taxon, but very few of them can or should be assigned a Linnaean rank.[20] Equally, one may argue, many of them should not be considered real groups or kinds.

To bring the question of ontological profligacy into sharper focus, consider three ontologies: according to the first, we ought to accept as real all phylogenetic species, and all monophyletic groups: call this the S+M ontology. According to the second, we ought to admit only species, not higher taxa, into our ontology: call this the SO ontology. According to the third, we ought to admit only monophyletic groups into our ontology (where ancestral species are ruled out as paraphyletic): call this the MO ontology. How do these ontologies score for ontological parsimony? Suppose a species *a* splits and gives rise to two species *b* and *c*, each of which splits and give rise to two species *d* and *e*, and *f* and *g*.

[19] I focus in this paper mainly on the real groups/natural kinds interpretation of the sufficiency thesis, but it should be noted that treating all monophyletic taxa as (objectively existing) concrete individuals (not kinds) may be an alternative way of elaborating the thesis. Some defenders of the species-as-individuals thesis have argued that monophyletic higher taxa are individuals in much the same sense, i.e. chunks of the genealogical nexus. See Boyd (1999) for a critique.

[20] I agree with the widespread (though not universal) view that cladism requires the abandonment of the Linnaean ranking system (Ereshefsky 1997; Griffiths 1994; Richards 2016: 153) and its replacement by an alternative. The Linnaean system, even in its greatly expanded modern form, doesn't contain anywhere near enough ranks for all the monophyletic taxa in the tree of life to be given a Linnaean rank.



According to S+M, there are ten real groups or kinds here: a*, b, c, d, e, f, g, b+d+e, c+f+g*, and *a+b+c+d+e+f+g*. According to SO there are seven real groups: the seven species. According to MO, there are seven real groups—*d, e, f, g, b+d+e, c+f+g*, and *a+b+c+d+e+f+g*. So SO and MO are equally parsimonious, but S+M is less parsimonious than both. Perhaps then, other things being equal, SO and MO should be preferred to S+O.

This may not be considered a very serious worry. The appeal to parsimony here may be questioned, and in any case parsimony considerations will only count against those who accept S+M, not those who accept MO (assuming that SO is the only serious alternative), and I have suggested that MO, not S+M, is the appropriate ontology for adherents of the cladist metaphysical principle.

A more telling concern may be that the sufficiency of monophyly position conflicts with the widespread view (including among cladists) mentioned above, that species are real in a way that higher taxa are not. According to the cladist metaphysical principle, all monophyletic taxa are equally real. So a monophyletic higher taxon is just as real as a monophyletic species. This runs contrary to the views expressed by at least some cladists (e.g. Eldredge and Cracraft 1980: 249) concerning the reality of species vis-à-vis higher taxa. Although the conflict between these views has not always been recognised, *if*[21] we think that species are real but higher taxa are not, we obviously have to reject monophyly as a sufficient condition of reality.[22]

[21] I am not endorsing this view, merely noting that if it is correct, it undermines the sufficiency view.

[22] It may be tempting to assimilate this view to the pragmatic view about classification I mentioned above: that the erection of higher taxa is purely a matter of convention or convenience, not answering to facts about nature. In the case of the cladists who hold the view (such as Eldredge and Cracraft), this would be a mistake. These theorists are *cladists* after all, meaning minimally that they accept the claim that classifying by phylogeny, and thus erecting monophyletic higher taxa, is a more objective and thus a superior approach to classification than rival

So there may be reasons for thinking that many monophyletic groups are not real groups or kinds, and thus that monophyly fails as a sufficient condition for reality. But further, once we have distinguished the classification and metaphysical questions, and noted that the (persuasive) arguments for the cladist classification principle do not obviously carry over to the cladist metaphysical principle, arguably we are left with few positive arguments *for* the view that all monophyletic groups are real. It is worth comparing the debate over the reality of species. A number of arguments have been offered for the view that species are objectively real, including the fact that anthropological evidence seems to show that many different kinds of human societies and cultures identify the same species taxa in nature (Atran 1999); the fact that species realism follows from certain well-supported macroevolutionary theses, such as Punctuated Equilibrium (Gould and Eldredge 1972, Gould 2002); and the fact (sometimes connected to the previous point) that species have a certain ontological status—they are concrete, cohesive, spatiotemporally bounded individuals, and thus are real, objective, discrete objects (Eldredge and Cracraft 1980).[23] Whatever we think about the cogency of such arguments, they do not appear to carry over to monophyletic groups in general, which should not be surprising, since the thrust of such arguments tends to be that species are ontologically special: they are real units or agents, in a way that higher taxa, whether monophyletic or not, are not (Mishler and Donoghue 1982: 491).

We can however interpret Griffiths' defence of the value of cladistic classifications (1994: 216–217) as an argument for the sufficiency view. Cladistic classifications (of both organisms and traits), he argues, are more informative than functional-adaptive classifications, because they are more predictively and explanatorily useful. If we know that a species belongs to a certain clade, we can predict more about its traits than we can on the basis on knowing that it occupies a certain ecological

approaches that make use of criteria other than ancestry. It is more plausible to interpret the view as a version of the one I am defending in this paper: classification by strict monophyly is objective, in that it respects real, objective divisions ('joints') in nature (the branching order of the tree of life), hence cladism with respect to classification is justified; however the higher taxa erected by such classifications (unlike species) may fall short of qualifying as real groups or kinds. Of course an alternative interpretation is that these thinkers are simply confused, not realising it is not coherent to endorse cladism while rejecting the reality of monophyletic higher taxa. I reject this interpretation since I do not regard this position as incoherent.

[23] We tend to be unreflective realists about particular, concrete, individual objects—tables, trees, horses, etc. So if species are, as the species-as-individuals (SAI) view claims, concrete, particular individuals, it may be hard to resist species-realism. Of course SAI may not be *necessary* for species-realism: one could hold that species are natural kinds (not individuals), for instance, and still be a species-realist. But SAI may still be *sufficient* for species-realism (or at least, strongly support it). Thank you to an anonymous referee for urging me to clarify the connection between SAI and species-realism.

niche. Species may share *some* superficial similarities with unrelated species that occupy the same niche, as a result of evolutionary convergence; but they share a great deal more similarities with other species in their clade, as a result of common ancestry. ‘Kiwis owe more of their characteristics to their descent from the common ancestor of birds (and, more recently, of the New Zealand rattites) than to adaptation to their current role as nocturnal, forest floor omnivores’ (216). Griffiths’ argument might be seen as supporting the view that all monophyletic clades are real kinds in the following way. Fundamentally, natural kinds support inductive inference, explanation, prediction, and generalisation. Knowing that an organism belongs to a particular monophyletic clade allows us to predict and explain a large number of its characters. Hence all clades are natural kinds (albeit of a historical nature).

One problem with the line of argument I’m attributing to Griffiths is that in its focus on relations of similarity, it may give too much ground to pheneticism. So Griffiths, following Fink, notes that crocs and birds share important, deep similarities (especially behavioural), as a result of their being closely related (216). The standard way of thinking about the relationship between similarity and phylogeny in this case is that on phenetic criteria, crocs would be grouped with lizards apart from birds (due to the divergence of the birds), while on phylogenetic criteria, crocs would be grouped with birds apart from lizards. But Griffiths is suggesting (I take it) that this is superficial: it may be that even on phenetic grounds of similarity of form and function (including behaviour), a good case could be made for grouping crocs with birds apart from the other ‘reptiles’. (Ridley (1986: 4–5) makes the parallel point about convergence: on a superficial interpretation, barnacles would be grouped with limpits apart from crabs (to whom they are more closely related as crustaceans) by pheneticists due to morphological convergence; but a closer study of the morphology of barnacles may well find that they more closely resemble crabs than limpits, such that the phenetic and cladistic classifications would agree with one another in this case.)

But the cladist holds that we should not go down the path of similarity at all: *even if* crocs share more similarities with lizards that they do with birds, they should still be grouped with birds because they share a more recent common ancestor. The similarity justification of cladistic classifications seems inherently risky, in its assumption that phylogenetic and phenetic classifications well tend to line up. What happens if they don’t? The whole motivation for cladism was precisely that arguments about similarity are irresolvable: no doubt crocs do share many interesting similarities with birds due to common descent (synapomorphies for the bird-croc clade); they also share many similarities with other ‘reptiles’ due to common descent (synapomorphies for the reptile-bird clade). Which set of similarities is more important for classification? Of course, in the context of crocs, birds and lizards, the shared characters of crocs and birds are synapomorphies, while the shared characters of crocs and lizards are symplesiomorphies. But this presupposes the cladist framework where the reconstruction of phylogeny is the goal. For pheneticists focusing on shared characters, with no interest in phylogeny, all shared characters are equivalent, and the question whether crocs are overall 'more similar to' lizards or birds may have no objective answer.

The second point I would make here is that there are challenges facing any attempt to vindicate the idea of clades as natural kinds in terms of the traditional notion of kinds as sets of similar entities that support induction and explanation, and are defined by an essence that explains why the members of the kind possess the features they do. As Griffiths notes, clades are fundamentally historical entities. If they are kinds defined by an essence, it would be a historical essence, presumably the clade's evolutionary origin in a common ancestor (Rieppel (2005) explicitly endorses this view). To count as the essence of a clade on the standard understanding of essences, this ancestry would need to be causally responsible for, and help to explain, the traits of the organisms in the clade. This would be the clade-level analogue of the historical essence view about species defended by Griffiths (1999), LaPorte (2004) and others. Even when applied to species, however, the historical essence view faces serious objections. Okasha, for example, has argued that an organism's ancestry (and indeed any other relational properties, such as ability to interbreed with other members of the species, that might be candidates for the species essence) does not cause, or help to explain, the organism's morphological traits (2002: 203–204).[24] '…the causal explanation of why an organism has the particular morphological traits it does will cite its genotype and its developmental environment … its belonging to (a particular chunk of the genealogical nexus) is not the explanation—or at least not the proximal explanation—of why it has the morphological traits that it does' (204). Okasha is not in fact rejecting the historical essence view: he thinks relational properties such as ancestry can count as species essences even if they don't cause or help to explain the traits of organisms. But most defenders of the historical essence view of species or clades do accept the traditional requirement that essences play this causal and explanatory role. If we accept the requirement, and if Okasha is right that a species' ancestry does not satisfy it, it follows that that ancestry cannot be the essence of a species. If that is true for species, it is just as true (if not more true) for clades.

A further problem for the historical essence account of clades is presented by Pedroso (2012; see also 2014). He notes that the main argument for the historical essence view is that it is required by cladism. That is, it follows from cladism that the essence of a biological taxon is its ancestry in the sense that if taxon X is the common ancestor of the

[24] See Nanay (2011) for a more metaphysical argument against the view that species essences cause and/or explain the features of organisms.

members of clade C in the actual world, then X is the common ancestor of the members of C in every possible world in which C exists. It is not possible to be a member of C and not to have descended from X. But, Pedroso argues, this does not follow from cladism. It is consistent with cladism that there are possible worlds in which C exists but its members do not have X as their most recent common ancestor. All that is required by cladism is that C be a monophyletic clade in every possible world. The common ancestor of the members of C can vary across worlds. Cladism entails only that the *cladogram* true of the clade in the actual world is true of the clade in all possible worlds. But of course the same cladogram is consistent with multiple incompatible phylogenetic *trees*, specifying different ancestors for the members of the clade. Thus it is not a necessary truth that some Y is a member of C just in case Y descends from X. Historical essentialism fails.

Pedroso, like Okasha, has presented serious problems for the historical essence approach to justifying the sufficiency thesis. In particular, in line with the argument of this paper, Pedroso has shown that one may be a cladist about classification without accepting (at least the historical essence version of) the sufficiency thesis. Of course, this does not show that there might not be ways of defending the sufficiency thesis other than that associated with the natural-kinds-defined-by-historical-essences view. But I am not aware of any plausible candidates.[25]

We have seen that one motivation for conflating the metaphysics and classification questions is the idea that both taxonomy and the metaphysics of kinds aim to 'carve nature at the joints'. There is

[25] It has been suggested by some theorists (e.g. Rieppel 2005) that the homeostatic property cluster (HPC) account of natural kinds associated with Richard Boyd (1991, 1999) (possibly in connection with the historical essence account) can be applied to taxa, including higher taxa, in a way that would justify the sufficiency (and possibly the necessity) thesis. Rieppel suggests that monophyletic taxa are HPC natural kinds (the sufficiency thesis), and that nonmonophyletic taxa are 'artificial' (the necessity thesis). I do not have the space to consider in detail HPC theory and its relation to monophyly; suffice it to say that it is questionable whether HPC theory is compatible with cladism. Ereshefsky (and his co-thinkers) have been arguing for a number of years that while cladism classifies by ancestry and genealogy irrespective of similarity, HPC kinds are ultimately similarity-based kinds, with the result that cladistic kinds will not always map onto HPC kinds (see Ereshefsky 2010, Ereshefsky and Matthen 2005, Ereshefsky and Reydon 2015). If that is correct, HPC theory will not be compatible with the sufficiency or necessity theses. Indeed this is how Boyd sees matters. He argues that some HPC kinds are paraphyletic and some are polyphyletic (2010: 693); and he suggests that to be a real kind it's not enough that a higher taxon be monophyletic—it has to satisfy other conditions as well (to do with his 'accommodation thesis'). Thus he rejects the necessity and sufficiency theses, though on different grounds from those presented here. Boyd's views about kinds, monophyly, and higher taxa are complex and subtle, and I can't hope to do justice to them here. But the following upshot of his argument seems in any case highly congenial to the line of reasoning I have been pursuing: 'We need not think of monophyletic groups as occupying some especially privileged … position relative to other natural kinds in evolutionary biology in order to insist that higher taxa must be monophyletic' (2010: 694).

no question that monophyly, and the objective order of evolutionary branching, represent real 'joints' in the natural world and its history. That there is an objective fact about the order of branching in the tree of life, and hence about evolutionary relationships, is the main argument supporting phylogenetic systematics (an argument I accept). But it does not follow, I have urged, that taxa formed on this basis are necessarily real groups or kinds. My rejection of monophyly as a sufficient condition for real groups or kinds can be understood as the claim that the kind of joint-carving that the construction of monophyletic taxa exemplifies is not of itself sufficient for carving up organisms into real groups or kinds. I do not claim that the cladist metaphysical principle does not carve at joints. I claim that carving at joints in the minimal sense is not sufficient for identifying real kinds. Other conditions must be satisfied.[26]

As an analogy, consider Kitcher's discussion of real kinds in astronomy (1992: 105). Kitcher, in his discussion of shifting 'reference potentials' of theoretical terms in science, highlights the ways in which the term 'planet' has shifted its reference throughout history; at one time referring to the known planets of our solar system excluding the earth; later referring to all the planets of our solar system including the earth; and finally referring to all the planets orbiting all the stars in the universe. He suggests that in the first case (reference to planets of our solar system excluding the earth) the term did not pick out a natural kind, but in the two subsequent cases it did. Thus, there is a natural kind comprising all and only the planets of our solar system. But this is, I'd suggest, implausible. There is certainly an objective 'joint' of a sort here—an objective division in nature—and 'planet' as referring to all and only the planets of our solar system carves at this joint. But despite this, many, I am assuming, would hesitate to regard the set of objects thus designated as a genuine natural kind (as opposed to the set of all planets of all stars, which has a stronger claim to constituting a natural kind). The predictive and explanatory value of the kind term 'planet' used in this restrictive sense is very limited indeed, and there are presumably no interesting laws true of all and only the objects picked out by it. Thus, a term can carve at a natural joint without picking out a natural kind.

[26] Here I follow Bird (2018), who distinguishes natural *divisions* in nature from natural *kinds*. He notes that green things are naturally similar to one another, such that there is a natural division of the world into green and non-green things, but green things do not form a natural kind. And he suggests we can imagine a world in which there are natural divisions but no natural kinds. One could thus be a weak realist about natural divisions without committing to the reality of natural kinds. I suggest that if our concepts correspond to natural divisions, they 'carve at the joints'. But only some natural divisions correspond to natural kinds. Thus, I claim that there is a natural division of the tree of life into monophyletic clades, but that this is not sufficient for those clades to count as natural kinds. Thank you to an anonymous referee for encouraging me to be clearer on this issue.

One may be tempted then to retreat to the claim that carving at the joints is necessary, but not sufficient, for picking out real kinds. That may be true, but I don't find any support here for the necessity of monophyly thesis, since there is no reason to suppose that monophyly is the only relevant 'joint' at which to carve up organisms into real kinds. If, say, 'predator', picks out a real kind then it carves at a joint: just not the ancestry joint (see below).

I conclude that since (a) there are no very convincing arguments for monophyly as a sufficient condition of reality, and (b) there are some good arguments against it, we should reject monophyly as a sufficient condition of reality. In the previous section I argued it is not necessary either. To paraphrase Dupre,[27] monophyly makes good sense for classification; it is something of a disaster for metaphysics.

4. *Beyond monophyly*

So monophyly does not appear to be the right criterion for determining the reality of groups and kinds. If not monophyly, what should be our criterion? A clue to this can be found by considering again Sterelny and Griffiths' suggestions about paraphyletic groups.

Sterelny and Griffiths' discussion makes clear the differing motivations for, and differing status of, the compromise classification view, and the compromise metaphysics view. With respect to the former, evolutionary taxonomists have wanted to allow paraphyletic taxa in large part because of morphological considerations. It is the great morphological dissimilarity of birds and reptiles, due to divergence,[28] that motivates the desire the keep Reptilia as a respectable higher taxon, and to elevate the birds to the same rank as the reptiles. And it is the inability of evolutionary taxonomy to consistently and non-arbitrarily apply this morphological criterion that ultimately undermines it, according to Sterelny and Griffiths and many others.

[27] 'Strict monophyly is an obvious desideratum from the point of view of mapping evolution. But from the point of view of classification it is something of a disaster' (2002: 431). Dupre, unlike me, is of course rejecting the cladist classification principle.

[28] In fairness, evolutionary taxonomists have not appealed merely to morphological criteria to justify their taxonomic decisions. They have also elaborated the concept of an adaptive 'grade': reptiles, mammals and birds are legitimate taxa of the same rank (traditional classes of chordates) because they possess different integrated adaptive complexes—they are each characterised by certain sets of adaptive innovations. Reptiles possess a certain suite of characters adapting them to a certain broad niche, as do mammals, and birds (Ridley 1986: 32-33; Brysse 2008: 305). So we have adaptive, not purely phenotypic, divergence and differentiation. The concept of adaptive grades has been criticised by cladists as being vulnerable to the same problems of subjectivity and arbitrariness as the purely morphological criteria (Ridley 1986, 33). Whether a putative taxon has evolved a sufficiently novel suite of adaptive innovations to count as a new 'grade' is not something that may be determined using objective criteria. The emphasis on adaptation in the notion of a grade has also been criticised by anti-adaptationists.

With respect to the metaphysical question on the other hand, the considerations are quite different. The reason Sterelny and Griffiths give for keeping paraphyletic groups is not primarily morphological. If the non-marine mammals constitute a real group or kind it is not primarily because the marine mammals have diverged morphologically from their non-marine ancestors and cousins, with the latter retaining a suite of features uniting them into a coherent higher taxon. Rather, it has to do with whether there are respectable evolutionary hypotheses about the non-marine mammals. It has to do with their role in evolutionary explanations. Thus we can frame an alternative *explanatory* criterion for the reality of groups or kinds: groups or kinds are real to the extent that positing them does important explanatory work for scientists.[29]

Devitt (2011) has defended a similar explanatory criterion for biological natural kinds, focusing on whether an entity's being a member of a putative kind is explanatory of the features of the entity. But he suggests that the question of realism—whether certain kinds exist objectively—has been conflated with the question of which kinds are *natural* kinds. '…the non-natural is being confusingly described as the non-real.' (165) On the realism question, reptiles obviously exist, he argues: the reptile kind is clearly a real kind that exists objectively. The interesting question is whether it is a *natural* kind: this depends on whether it is an *explanatorily significant* kind. Being a reptile may, he says, be like being a cousin: cousins exist, but being a cousin is not explanatorily significant. Against Devitt, I agree with Griffiths and others that *if* 'reptile' does not name a natural (explanatory) kind, then it does not name a kind at all: there is no reptile kind and reptiles do not exist. 'Reptile' would be non-referring. The appropriate analogy is not with 'cousin', but with 'witch', or 'phlogiston'. The latter are posits of false theories: when we reject the theories, we reject the existence of the kinds posited by the theories, and declare the putative kind terms non-referring. 'Witch' does not refer to a non-explanatory but real kind—it doesn't refer to a kind at all (which is not to say, of course, that the women who this term was applied to did not exist); 'phlogiston' does not refer to a non-explanatory substance—it doesn't refer to a substance at all. 'Reptile' is theory-laden in just the way that 'witch' and 'phlogiston' are; if the theories that treat the reptile kind

[29] This is a version of the Quinean explanatory criterion for ontology, which says that we should be ontologically committed to the entities the positing of which is required for our best scientific (and perhaps philosophical) explanations, or those that enhance the explanatory power of our theories. As in the literature on the broader Quinean criterion, the notions of 'explanation', 'explanatory power', and 'best' in 'best explanations' will here be assumed to be sufficiently intuitively clear. But for a summary of different accounts of the nature of scientific explanation see Woodward and Ross (2021); and for a useful discussion of what makes for a good inference-to-the-best-explanation, in particular in biology, see Lewens (2007: ch. 4). Thank you to an anonymous referee for suggesting I clarify this point.

as an explanatorily significant natural kind are false, there are no reptiles.[30] Thus, while I agree with Devitt about the explanatory criterion for natural kinds, unlike Devitt I take this to be a criterion for reality, not just naturalness.

On this way of looking at it, Sterelny and Griffiths' sympathy for the compromise metaphysical view, but lack of sympathy for the compromise classification view, is intelligible. While the latter involves a 'mixed' criterion that attempts to do justice to both similarity and relatedness in classification, and as such cannot avoid subjectivity and arbitrariness with respect to the aims of classification, the former is an application of a quite straightforward explanatory criterion of naturalness and reality. The criteria have different statuses, so it is not surprising that the compromise paraphyletic-friendly positions they give rise to inherit these different statuses.[31]

This is relevant when considering the following natural response to my view. If we think there are paraphyletic real groups or kinds, then why not allow paraphyletic higher taxa corresponding to those kinds? Conversely, if we are rejecting paraphyletic taxa, how can we allow paraphyletic real groups? The response to this is that there are persuasive arguments against allowing paraphyletic taxa, but that these don't carry over to paraphyletic real groups/kinds. As we have seen, paraphyletic higher taxa (at least, in the evolutionary taxonomy tradition) can only be justified on phenetic grounds of similarity and dissimilarity.[32] And such grounds do not provide for objective classifications. Only strict monophyly, corresponding to the objective order of branching of the tree of life, can provide objective classifications.[33] But

[30] I do however agree with Devitt that it is not obvious that such theories are false, i.e. not obvious that paraphyletic kinds such as Reptilia are not explanatorily significant kinds.

[31] Another way of putting this is that the evolutionary taxonomy position on classification involves a compromise with phenetics, whereas the explanatory argument for allowing paraphyletic groups does not. (So really it's wrong to call the compromise metaphysical view a *compromise* view.)

[32] Paraphyletic ancestral species are a somewhat different case.

[33] That paraphyletic taxa should be rejected is common ground among cladists, but there has not always been sufficient clarity about *why* they should be rejected. For instance, Eldredge and Cracraft (1980) argue that the problem with paraphyletic taxa is that they are 'not-A' groups, i.e. groups defined by the *lack* of some property or set of properties. They suggest (a) that not-A groups are less natural than A groups (defined by possession of positive properties); (b) that eliminating them has been important in making progress in systematics; and (c) that cladism is the natural culmination of this tendency. Only A groups. i.e. monophyletic groups, should be allowed in a classification. The problem with this is that A groups are defined phenetically: by possession of certain defining (essential) properties. If cladistic groups are not *defined* by (but rather are identified using) synapomorphies (Ridley 1986) then being monophyletic is not sufficient for being an A group (one whose members all possess the defining property); and if A groups can be phenetic (not phylogenetic) groups, then being monophyletic is also not necessary for being an A group. Eldredge and Cracraft suggest (164) that reptiles, fish etc. are illegitimate

this argument does not apply to the paraphyletic real groups/kinds my analysis allows, because these are not identified using phenetic criteria. They are identified using explanatory criteria.

The explanatory criterion is more general than the criteria that define the different schools of classification, in that it says nothing about either similarity or evolutionary relatedness. It says that we should be ontologically committed to all and only those groups that feature in in well-confirmed scientific explanations and hypotheses. It does not say 'and these must be groups of organisms that form a coherent evolutionary unit (i.e. are monophyletic)' or 'and these must be groups of organisms that are united by similarity'. No doubt often the groups that satisfy the former, general, condition will also satisfy one or both of the more specific conditions. But they need not. Groups that satisfy the general condition may not be defined by relations of similarity, and may not be monophyletic.

Arguably they need not even be paraphyletic. There is no reason in principle why the explanatory criterion could not certify the reality of some polyphyletic groups. It is generally agreed that polyphyletic groups defined phenetically (merely in terms of shared characters)—creatures with wings; creatures with eyes etc.—are not legitimate taxa, and do not form real kinds. That they don't form real kinds is supported by the explanatory criterion: there are no interesting biological hypotheses concerning these 'kinds'. They play no role in biological explanations. Other polyphyletic groups may have a greater claim to being real kinds however (even if they are not legitimate taxa in the context of systematics), for example ecological kinds such as 'predator' (Wilson et al. 2007: 194–5) or 'parasite'. In ecology 'predator' has real biological significance, appears to play an essential role in ecological explanations, and so on.[34] It plausibly count as a real kind on the ex-

because they are not-A groups. But this just shouldn't be the issue from a cladist point of view: even if they *were* defined by a particular (positive) property or set of properties, so counted as A-groups, they would still be illegitimate because paraphyletic. The cladist ought to insist that the whole question of possession of (intrinsic) properties, and thus the issue of positive vs. negative properties, is a red herring. The sole issue for classification is common ancestry and monophyly. It is this that makes birds and mammals, but not fish and reptiles, legitimate taxa, not any issue to do with A vs. not-A groups.

[34] '...biologists see [categories such as 'predator'] as corresponding to kinds because of their explanatory and predictive value. Individual predators are predators not in virtue of being integrated parts in a larger individual, but in virtue of certain intrinsic and relational properties that they tend to share and which underwrite certain explanations, predictions, and generalisations...' (Wilson *et al* 2007: 195. See also Devitt 2011. However see Griffiths (1994) for reasons to be sceptical about the value of 'purely' functional/ecological categories such as 'predator'. All useful functional categories, he suggests, are historically constrained, and historically constrained functional kinds can be paraphyletic, but not polyphyletic; 218). Wilson *et al* are here arguing that some real kinds in biology are not individuals, but their point also supports my claim that some real kinds of organisms in biology are not

planatory approach.[35] Griffiths notes that generalisations about such ecological kinds occur at the functional-adaptive level of biological explanation, in which organisms and traits are classified in terms of their adaptive or ecological role (Griffiths 1994: 215–217). Such (abstract) functional roles are multiply realised by underlying cladistic kinds (the kind 'predator' is realised by many different lineages within different clades).[36] On this picture there are polyphyletic real kinds, identified at the functional-adaptive level, but the taxa that realise those kinds are monophyletic clades, identified at the historical-phylogenetic level.[37]

Thus, as with the above paraphyletic examples, on the explanatory criterion such putative ecological (polyphyletic) kinds are not ruled out simply by virtue of being non-monophyletic. It is a major virtue of that approach that it is flexible enough to potentially accommodate a wide range of biological kinds quantified over by workers in different areas within biological science.

Here the difference between the classification question and the metaphysical question is especially clear. The explanatory argument for admitting polyphyletic kinds such as 'predator' is not at all impugned by the widely accepted (even by evolutionary taxonomists) and persuasive arguments for rejecting polyphyletic taxa. There is clearly no predator taxon, but it is plausible that there is a predator real kind.

The explanatory criterion provides, I suggest, a sounder criterion for identifying real groups or kinds of organisms than does the cladist metaphysical principle of the necessity and sufficiency of monophyly. Of course, these may not *necessarily* have been in conflict: it might have turned out that in applying the explanatory criterion, the necessity and sufficiency principles were vindicated. Indeed this is likely to be the response from supporters of the cladist metaphysical principle to my opposing to it the explanatory criterion—that these are not in competition, that rather the cladist principle is *justified by* the prior and

monophyletic (given the close relationship between cladism and the species-as-individuals thesis, this should not be surprising).

[35] In this respect Sterelny and Griffiths' defence of the reality of paraphyletic groups proves too much (for their liking). They appear to want to allow paraphyletic but not polyphyletic kinds. But the criterion they appeal to—whether there are interesting biological hypotheses about the group in question—would appear, as we have seen, to certify the reality of at least some polyphyletic groups. That is, the compromise metaphysical view ((5) above) is unstable. Once you recognise paraphyletic groups on those grounds, you also have to recognise polyphyletic groups. There is no argument *of this sort* to show paraphyletic groups can be admitted that does not also show polyphyletic groups can be admitted.

[36] Griffiths presents this two-level picture and acknowledges its attractiveness but goes on to criticise it somewhat later in the paper.

[37] Consider, as another example, Hull's suggestion (1988: 215) that 'cosmopolitan species' is a candidate for a natural kind that may feature in laws of nature, presumably by virtue of its explanatory credentials. If this is a real kind it is a polyphyletic one that is realised by cladistic taxa (i.e. species) but is not itself a taxon.

more general explanatory principle (or something like it). But it should be clear why I hold that in fact they do conflict. It is very plausible that positing paraphyletic ancestral species, and polyphyletic ecological kinds, is explanatorily valuable. And I have challenged the claim, implicit in Griffiths and others' work, that all monophyletic groups are explanatorily significant kinds. The explanatory principle undermines, rather than supports, the necessity and the sufficiency theses.

It is important to see, firstly, that the explanatory criterion I have proposed is a criterion for determining only which groups of organisms are real, it is not intended to be an account of the nature of natural kinds in biology generally, much less a theory of natural kinds in general. Secondly, it is intended to be a *criterion* for determining whether certain groups of organisms are real groups or kinds, not a complete account of the metaphysics or epistemology of these kinds, or natural kinds in general. There has been much philosophical work recently devoted to the question of whether or not natural kinds are mind-independent, whether they should be defined in metaphysical or epistemic terms, and if the former, what their metaphysical status is—whether they are reducible to sets, universals, or something else, or are *sui generis* entities (for important recent contributions see Bird 2018 and Franklin-Hall 2015).

These are interesting and important questions but I do not need to take a stand on them. In particular, the fact that their role in biological explanations is our criterion, that is, best (perhaps only) evidence, for the reality or naturalness of groups of organisms does not entail that their explanatory role or value is *constitutive* of their naturalness, in a way that would suggest an anti-realist or epistemic account of natural kinds, such as those defended by (on some interpretations) Boyd (1991, 1999), Magnus (2012, 2014) and Ereshefsky and Reydon (2015). Groups of organisms may *be* natural kinds in virtue of entirely mind-independent facts, yet it might still be the case that we only *know* they are natural kinds in virtue of their role in scientific explanations. As far as I can see the explanatory criterion I have defended is consistent with all (or at least most) of these more abstract theories of the fundamental nature of natural kinds.

Conclusion

It has been assumed that if one accepts cladism with respect to classification, one must accept what I have called the cladist metaphysical thesis, the claim that all and only real groups or kinds of organisms are monophyletic. In section 2 I argued that this is not the case, that the classification and metaphysics questions are logically distinct, such that cladists (with respect to classification) *can* reject the cladist metaphysical thesis. In section 3 I argued that the cladist metaphysical thesis is implausible: there are real groups or kinds that are not monophyletic, and plausibly monophyletic groups that are not real or natural.

Thus cladists with respect to classification (and others) *should* reject the cladist metaphysical thesis. In section 4 I explicitly endorsed an alternative and superior explanatory criterion for the reality of groups or kinds (implicit in my earlier criticisms of the cladist metaphysical thesis). This need not amount however to a rejection of cladism in general, so long the metaphysical question is sharply distinguished from the question of classification. Cladistic classification may survive the rejection of cladistic metaphysics.

References

Ashlock, P. D. 1971. "Monophyly and Associated Terms." *Systematic Zoology* 20 (1): 63–69.

Atran, S. 1999. *Folkbiology*. Cambridge: MIT Press.

Baum, D. A. and M. J. Donoghue. 1995. "Choosing among Alternative 'Phylogenetic' Species Concepts." *Systematic Botany* 20 (4): 560–573.

Bird, A. 2018. "The Metaphysics of Natural Kinds." *Synthese* 195: 1397–1426.

Boyd, R. 1991. "Realism, Anti-foundationalism, and the Enthusiasm for Natural Kinds." *Philosophical Studies* 61: 127–148.

____ 1999. "Homeostasis, Species, and Higher Taxa." In R. A. Wilson (ed.). *Species: New Interdisciplinary Essays*. Cambridge: MIT Press, 141–185.

____ 2010. "Homeostasis, Higher Taxa, and Monophyly." *Philosophy of Science* 77 (5): 686–701.

Brysse, K. 2008. "From Weird Wonders to Stem Lineages: The Second Reclassification of the Burgess Shale Fauna." *Stud. Hist. Phil. Biol. & Biomed. Sci* 39: 298–313.

Christofferson, L. 1995. "Cladistic Taxonomy, Phylogenetic Systematics, and Evolutionary Ranking." *Systematic Biology* 44 (3): 440–454.

Cracraft, J. 1981. "Pattern and Process in Paleobiology: The Role of Cladistic Analysis in Systematic Paleontology." *Paleobiology* 7 (4): 456–468.

Devitt, M. 2011. "Natural Kinds and Biological Realisms." In J. K Campbell, M. O'Rourke, M. H. Slater (eds.). *Carving Nature at its Joints: Natural Kinds in Metaphysics and Science*. Cambridge: MIT Press, 155–174.

Donoghue, M. J. 1985. "A Critique of the Biological Species Concept and Recommendations for a Phylogenetic Alternative." *The Bryologist* 88 (3): 172–181.

Dupre, J. 1981. "Natural kinds and Biological Taxa." *The Philosophical Review* (1): 66–90.

____ 1993. *The Disorder of Things: Metaphysical Foundations for the Disunity of Science*. Cambridge: Harvard University Press.

____ 2002. "Hidden Treasures in the Linnean Hierarchy." *Biology and Philosophy* 17: 422–433.

Eldredge, N. and J. Cracraft. 1980. *Phylogenetic Patterns and the Evolutionary Process*. New York: Columbia University Press.

Ereshefsky, M. 1997. "The Evolution of the Linnaean Hierarchy." *Biology and Philosophy* 12 (4): 493–519.

____ 1998. "Species Pluralism and Anti-Realism." *Philosophy of Science* 65: 103–120.

____ 2010. “Species.” In E. N. Zalta (ed.). *Stanford Encyclopedia of Philosophy*. Accessed 21 Sept. 2011 <http://plato.stanford.edu/entries/species/>

Ereshefsky, M. and M. Matthen. 2005. “Taxonomy, Polymorphism, and History: An Introduction to Population Structure Theory.” *Philosophy of Science* 72 (1): 1–21.

Ereshefsky, M. and T. A. C. Reydon. 2015. “Scientific Kinds.” *Philosophical Studies* 172: 969–986.

Franklin-Hall, L. 2015. “Natural Kinds as Categorical Bottlenecks.” *Philosophical Studies* 172 (4): 925–948.

Gould, S. J. 2002. *The Structure of Evolutionary Theory*. Cambridge: Harvard University Press.

Gould, S. J. and N. Eldredge. 1972. “Punctuated Equilibria: an Alternative to Phyletic Gradualism.” In T. J. M. Schopf (ed.). *Models in Paleobiology*. San Francisco: Freeman, Cooper & Co., 82–115.

Griffiths, P. E. 1994. “Cladistic Classification and Functional Explanation.” *Philosophy of Science* 61: 206–227.

____ 1999. “Squaring the Circle: Natural Kinds with Historical Essences.” R. A. Wilson (ed.). *Species: New Interdisciplinary Essays*. Cambridge: MIT Press, 209–228.

Hull, D. 1979. “The Limits of Cladism.” *Systematic Zoology* 28 (4): 416–440.

Kitcher, P. 1993. *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford: Oxford University Press.

LaPorte, J. 2004. *Natural Kinds and Conceptual Change*. Cambridge: Cambridge University Press.

Lewens, T. 2007. *Darwin*. New York: Routledge.

Magnus, P. D. 2012. *Scientific Enquiry and Natural Kinds: From Planets to Mallards*. Palgrave Macmillan.

____ 2014. “NK≠ HPC.” *The Philosophical Quarterly* 64: 471–477.

Mayr, E. 1942. *Systematics and the Origin of Species*. Columbia University Press

Mishler, B. and M. J. Donoghue. 1982. “Species Concepts: A Case for Pluralism.” *Systematic Zoology* 31 (4): 491–503.

Mishler, B. and R. Brandon. 1987. “Individuality, Pluralism and the Phylogenetic Species Concept.” *Biology and Philosophy* 2: 106–123.

Nanay, B. 2011. “Three Ways of Resisting Essentialism about Natural Kinds.” In J. K Campbell, M. O’Rourke, M. H. Slater (eds.). *Carving Nature at its Joints: Natural Kinds in Metaphysics and Science*. Cambridge: MIT Press.

Okasha, S. 2002. “Darwinian Metaphysics: Species and the Question of Essentialism.” *Synthese* 131 (2): 191–213.

Pedroso, M. 2012. “Essentialism, History and Biological Taxa.” *Stud. Hist. Phil. Biol. & Biomed. Sci* 43: 182–190.

____ 2014. “Origin Essentialism in Biology.” *The Philosophical Quarterly* 64 (254): 60–81.

Podani, J. 2010. “Monophyly and paraphyly: A discourse without end?” *Taxon* 59 (4): 1011–1015.

Quinn, A. 2017. “When is a Cladist not a Cladist?” *Biology and Philosophy* 32: 581–598.

Richards, R. A. 2016. *Biological Classification: A Philosophical Introduction*. New York: Cambridge University Press.

Ridley, M. 1986. *Evolution and Classification: The Reformation of Cladism*. London: Longman.

____ 1989. "The Cladistic Solution to the Species Problem." *Biology and Philosophy* 4: 1–16.

____ 1993. *Evolution*. Cambridge: Blackwell.

Rieppel, O. 2005. "Monophyly, Paraphyly, and Natural Kinds." *Biology and Philosophy* 20: 465–487.

Simpson, G. 1961. *The Principles of Animal Taxonomy*. Columbia University Press.

Sober, E. 1988. *Reconstructing the Past: Parsimony, Evolution and Inference*. Cambridge: MIT Press.

____ 2000. *Philosophy of Biology*. 2nd ed. Oxford: Westview Press.

Sterelny, K. and P. Griffiths. 1999. *Sex and Death: An Introduction to Philosophy of Biology*. Chicago: The University of Chicago Press.

Wiley, E. O. 1992. "The Evolutionary Species Concept Reconsidered." M. Ereshefsky (ed.). *The Units of Evolution: Essays on the Nature of Species*. Cambridge: MIT Press, 81–92.

Wilkins, J. 2009. *Species: A History of the Idea*. Berkeley: University of California Press.

Wilson, R. A., M. J. Barker and I. Brigandt. 2007. "When Traditional Essentialism Fails: Biological Natural Kinds." *Philosophical Topics* 35 (1 & 2): 189–215.

Woodward, J. and Ross, L. 2021. "Scientific Explanation." *Stanford Encyclopedia of Philosophy*. Stanford University Centre for the Study of Language and Information. <https://plato.stanford.edu/entries/scientific-explanation/>

Croatian Journal of Philosophy
Vol. XXII, No. 64, 2022
https://doi.org/10.52685/cjp.22.64.4
Received: October 25, 2021
Accepted: January 20, 2022

Identity of Dynamic Meanings

PAVEL ARAZIM[*]
Czech Academy of Sciences, Institute of Philosophy, Prague, Czech Republic

Inferentialism has brought important insights into the nature of meanings. It breaks with the representationalist tradition that sees meanings as constituted primarily by representing some extra-linguistic reality. Yet the break with tradition should be pursued further. Inferentialists still regard meanings as static, and they still do not entirely abandon the idea of fully determined meaning. Following Davidon's ideas about meanings as constituted only in the course of a specific conversation, I propose a dynamic account of what meanings are. They are described as entities belonging to the dynamic realm of Henri Bergson's duration. The inhabitants of this realm live in constant movement and development which is more essential to them than the stages that this development goes through. My account brings about a rejection of the notion of strict literal meaning and therewith also of the contrasting notions such as ambiguity. Meaning is understood as a dynamic entity that is characterized rather by its history than by its nature.

Keywords: Meaning; identity; development; rule; inferentialism; Bergson.

1. Introduction

The notions of ambiguity and vagueness belong to the usual conceptual toolkit of linguists. They surely have their justification in the usage the linguists make of them, yet they bear an understanding of linguistic meaning which I believe poses some important problems. I will indicate how our understanding of what meaning is should be modified and what understanding of vagueness and ambiguity it will bring about.

[*] This article was supported by grant 20-18675S *The Nature of Logical Forms and Modern Logic*, led by Vladimír Svoboda from the Institute of Philosophy of Czech Academy of Sciences.

Our concern will ultimately be with the identity criteria of meaningful expressions.

Both ambiguity and vagueness are of importance also for the philosophy of language, which is documented by the attention these phenomena have been paid to in philosophical literature such as Williamson(1994), Keefe (2000) and Smith (2008). An account of ambiguity and vagueness determines what understanding one has of linguistic meaning and of language. In particular, it determines the identity criteria of meaning, i.e. the question of how you understand what the boundaries of meaning are. Where does one meaning end and another begin?

How are ambiguity and vagueness usually understood? Let us review them in order. Ambiguity means that a given expression has more than one meaning. This can obviously lead to confusion, as it can sometimes be problematic to decide which meanings are meant in a given context. Even worse, such a misunderstanding can be abused by manipulators who switch between the various meanings in the course of an argument and thus beguile the audience. If the possible confusion is considered as particularly dangerous, an obvious remedy is to disambiguate which means to keep only one of the meanings associated with the given expression and, if it is requisite, reserve different expressions for the other meanings.

Vagueness is closely related to ambiguity. A vague expression has a strongly context-dependent meaning. The adjective *high* is typically considered as vague, as it points to a different height when we speak about elephants than when we speak about rabbits. A vague expression can behave in a manner similar to that of ambiguous expressions and can pose similar threats. Nevertheless, vagueness also has many specifics which would complicate my argument too much. Therefore, I will limit my attention to ambiguity, as it is important enough.

Ambiguity can be evaluated from many perspectives. It would be a great exaggeration to claim that it is generally seen as defective. Linguists certainly do also investigate the positive aspects of these phenomena. But still, there is a tradition, quite characteristic of analytic philosophy, to regard ambiguity as problematic. Think of the widely shared ideal of Carnapian explication, as it plays a role already in Carnap (1928). Other things being equal, it is considered more or less by default as progress when a common expression is replaced by less ambiguous one. Or, in the best-case scenario, all the ambiguity vanishes. Such a view presupposes quite a strong notion of an identity of a given expression.

I will focus on the inferentialist understanding of what a meaning consists of. I will argue that this account presupposes the possibility of a fully determinate meaning in its usual understanding. I argue that this, nevertheless, is a confused idea, as meaning always leaves something open. My thesis is stronger than contextualism or pluralivaluationism which presuppose that meaning can be determined by context.

My thesis is more radical, namely that the meaning is indeterministic and the contexts in which it enters cannot be fully specified in advance. I illustrate my idea of the dynamic nature of meaning by exploiting the ideas about a dynamic reality by Henri Bergson. This changes the understanding of ambiguity and of the identity of meaning. Although my discussion is primarily focused on meaning as understood by inferentialists, it is purported to confound other accounts which presuppose the idea of a determinate meaning. Nevertheless, there is still so much I endorse in inferentialism that my view can be still seen as a variety thereof. My position could be called *dynamic inferentialism*. We begin by considering what constitutes meaning in the first place. At least for the inferentialists.

2. *Inferential relationships*

Certainly, lots of theories trying to explain what meaning consists of have been proposed, and I cannot hope to consider all of them and then choose the best. I will focus on one which I believe is particularly strong and enables an illuminating view of the identity of a given expression. I choose this view because I think it can be modified in a fruitful way to suit what I want to say about meaning here.

The approach I will start with here is inferentialism, as it was hinted at by Wittgenstein and then subsequently formulated as a doctrine by Sellars, Brandom and Peregrin.[1] What is meaning according to the inferentialists? In the first place, the meaning of a sentence is explained by the inference relations it is featured in. These function according to rules[2] that specify what can and cannot be inferred from a given set of propositions. The meaning of a sentence is constituted by what it follows from and what follows from it, possibly with further premises. Thus, the meaning of a sentence such as *Rex is a dog* is constituted by such relations as those which tell us that we can infer it from *Rex is a dachshund* and we can infer *Rex is a mammal* from it. These inferences, then, are correct due to certain general rules. For example, that every dachshund is a dog and that every dog is a mammal.

This account has been subjected to many discussions since Brandom presented it.[3] I find it quite satisfactory, yet I will not spend time going back to the old controversies, particularly about the worries as to

[1] The most relevant sources are Wittgenstein (1953), Sellars (1974), Brandom (1994) and Peregrin (2014).

[2] I should note that when speaking about *rules,* I primarily mean inference rules in the whole article. Nevertheless, it is not just for brevity that I speak of *rules* more than of *inference rules*. A deeper reason is that I consider inference rules which constitute our language as intelligible only in the context of many other rules, in the spirit of paragraph 7 of Wittgenstein(1953):´I shall also call the whole, constisting of language and the actions into which it is woven, the *language game.*´

[3] Brandom was attacked in Lepore (2007), replies to the criticism can be read for example in Peregrin (2014).

whether it might not be all too idealistic, as it makes too much depend on the rules which we institute. It has great advantages, particularly in showing how meaning is not something enshrined in our mutually inaccessible minds or platonic heaven. But here, I will focus mainly on the fact that it particularly well represents what the idea of a completely defined meaning could be. By this I mean a meaning without ambiguity. The idea is simply that of having the inference rules specified, saying exactly what sets of premises the given sentence is a consequence of and exactly which consequences it has with which further premises.

Peregrin (2014: 50) expresses the inferentialist notion of the meaning of a sentence with particular clarity and technical precision. I will not reproduce his definition in detail but the basic idea is that for any sentence A, one can determine both what it is inferable from and what can be inferred from it, possibly with further premises. So, on the one hand, there is the set S of all sets of sentences from which A can be inferred. On the other hand, for any set of premises P, what follows from P and A together is specified. Taking these two ingredients together, we have a specification of inferential behavior of A, denoted as the *inferential potential* of A, abbreviated as IP(A). And for inferentialism, this means specifying the meaning of the sentence A. One can call this the proposition which is expressed by A. Regarding subsentential expressions, they are defined by their contributions to the meanings of sentences. If we call the meanings of subsentential expressions in a slightly idiosyncratic manner *concept,* then we can say that propositions determine concepts. The basic idea is that we first get acquainted with a limited number of sentences and their inferential relations and then, using substitutions extract meanings of subsentential expressions from them, enabling us to compose a potentially unlimited number of new sentences. This is described in Peregrin (2014: 62), based on Quine (1960: 9). This means that it is legitimate in my framework to speak indifferently of meaning and cover both the meanings of sentences and words or more complex subsentential expressions, i.e., cover both propositions and concepts.

IP(A) thus formally represents what the meaning of a sentence A is for an inferentialist. When we strive for a Carnapian explication of what meaning is for inferentialists, I think IP is as good as we can get. From this perspective, I do not want to replace it by an alternative definition. I would rather want to explain why we should also look for a different kind of understanding besides Carnapian explication. My approach will underline the dynamic aspect of inference practices and therewith of meaning. I will try to show what IP could prevent us from appreciating. I can begin by noting that IP is obviously a great idealization. No single speaker of a given language can overview all the possible inferences a given sentence can feature in. Anyone can thus have access at best only to a part of IP.

But there are more reasons to become suspicious. Let us return to dogs. How exactly is the concept of a dog specified? Should I infer from Rex's being a dog that he also has lungs? What if we discover creatures that are completely like dogs yet lack lungs? Should we infer that we have discovered a new kind of dog? If we are uncertain about this particular inference, then we should look up the general rules and see if the rule that all dogs have lungs holds. Yet we are just as uncertain as with the particular case of Rex. The rule that would decide our dispute is not available in the best list of rules that we have at our disposal. Does it mean that we cannot know what the correct answer is? Such an approach would be absurd. Rather than us being ignorant about the truth of the matter, there is simply no truth to the matter. At least not yet. Of course, we can make a decree one way or another. We can decide to regard this inference as valid or as invalid. Such a decree, if accepted, will then be normative for further usage. That is, the decree will be normative if it is successful.

We see that something in the meaning of the expression *dog* and therewith also of the related expressions such as *dachshund* and *mammal* was previously not under our control. It was not explicitly stated whether dogs must have lungs and the most adequate thing to say would be that it was objectively undetermined which answer was correct. Yet even this answer is doubtful, as the actual usage might have tended to move these expressions into one of the two possible directions. In this way, the expression is not in our control and the rules have to be rendered explicit in order to get the expressions more under our control. What is not explicit remains in a shadow and possibly indefinite. The notions of being merely implicit and of being indefinite are thus closely related, even if they are not the same. The relation I am hinting at should be clearer by the end of the article when we will understand how rules work in more detail, in particular how they emerge from and interact with our normative attitudes.

But what we have seen illustrated in one suggestive example about dogs and lungs can be generalized into more systematic reasons for believing that meaning cannot generally be in this manner explicit and therewith definite. Let us get acquainted with these reasons.

3. *Arguments against definite meanings*

We will present a global and two more local arguments against the notion of definite meanings.

3.1 *Global argument – the circularity argument*

This argument can be traced back to Wittgenstein and is reiterated, among others, by Brandom. The basic idea is that if you define the meaning of a given expression, you rely on your understanding of the expressions used in that very definition. For instance, when you define, say the expression A, then you use the expressions B, C and D to do so. These expressions must be clear as they stand. Should they be themselves unclear, we can, of course, continue by defining them by means of E, G and H and the process can go on. Yet if we never stop, then we sooner or later have to use A or some other expression from this succession A, B, C, ... anew and then we are obviously in trouble. This means that an ideal of a fully defined expression is indeed illusory.

This, however, does not entail that requirements for clarity are illegitimate and that we cannot criticize somebody for using an insufficiently defined expression. Only that the precision is always relative to a given context and can later always be found in some way partial. This might lead one to a Davidsonian view about how meaning is constituted only in a specific dialogical situation. We will return to this topic later to see to what degree we can embrace Davidson's position.

3.2 *Local argument number one*

A further argument for there being no entirely definite meanings is local in that it does not need to operate with the perspective of the whole of language or its vocabulary. It rather just focuses on the given expression and those most closely related to it, although all expressions are interrelated to some degree. Let us abstract from the fact that we always have finite equipment of possible explainers as we speak of a language with a finite vocabulary. Perhaps we can go on defining the expression A mentioned in the previous section by always new expressions. So we go beyond B, C, D, E, G, H to I and so on. Why cannot this work? The scenario with the sing-post due to Wittgenstein[4] will help us see where the problem lies.

Do we understand what a signpost instructs us to do? Well, typically we do, yet maybe we can in some contexts start doubting. Then we might get an explanation, perhaps that it is the sharp end of the arrow-shape that points in the direction we should go. But then again, we might want to get an explanation of this explanation. Obviously enough, this process would then continue, and we would embark on an infinite regress. Again, this does not mean that a request for an explanation is illegitimate. Only that there are some limits to it, in the given context it is only up to a certain stage that a request for further explanation is meaningful. As Wittgenstein (in fact, already Aristotle) also puts it, every explanation has to stop at some point. It has to stop in order to be an explanation at all.

How do we recognize this point? That is in general very difficult to tell. An answer which would suggest itself would be a point at which the explanation is already self-evident. Such an account is in a way true but needs to be specified further; otherwise, it can be more misleading than illuminating. The misleading impression that is not easy

[4] The famous sign-post is featured in aphorism 85 of Wittgenstein (1953).

to eliminate is that the self-evident has to be such in all contexts. Yet as I read him, Wittgenstein shows us that anything can be questioned and doubted in an appropriate context. Just think of the signpost, the explanation of which may itself require an explanation.

And furthermore, think of the example with dogs and lungs, Wittgenstein's ideas on number series and of quaddition of Kripke (1982).[5] The self-evidence is therefore itself only relative to a given context. Wittgenstein shows us that the doubt stops making sense at a given point. It becomes unclear whether the person who pretends to raise the new doubt understands the expression she uses. Further explanation is not possible at some points but that does not mean that these are the points at which all indeterminacy has been eradicated. This is because the expressions that might come close to being self-evident in these contexts quickly enter new contexts where they lose this status. They prove more interesting than they seem.

3.3 *Local argument number two – new contexts*

The last argument I offer is maybe a little bit less ingenious and more straightforward but its straightforwardness leads us directly to the particular point I want to make about meaning. The point is simple – it is the very essence of any expression or concept to adapt to various new and unprecedented contexts it enters into. Every context opens up new questions and indeterminacies to which the concept has to react and develop correspondingly. Whether all dogs must have lungs usually is irrelevant and therefore undecided, yet in some situations it may well become the key question, so we have to decide and adapt our original concept in a reasonable way.

The idea of a perfectly explicit and determinate expression, all the questions of the meaning of which are decided one way or another, is also an idea of an expression that is isolated from all the contexts. If we do not want to downplay the real influence of new contexts, we have to consider them as genuinely new. This means that the rules of the given language do not in advance establish how these contexts should be accommodated. Of course, many contexts are in various ways analogous

[5] The problem of continuing the number series is presented in paragraph 185 of Wittgenstein (1953), while the quaddition problem is introduced in Kripke (1982). Wittgenstein notes that even the most simple number series such as '2, 4, 6, 8, …' can be continued in countless ways. Besides the naturally looking continuation ´10, 12, 14, ...', one can also think, among many others, of '2, 4, 6, 8' so that we continue reiterating the quadruple. As for Kripke, he noted that everyone has learned the concept of addition by attending to a finite number of specific additions. Therefore, for everyone, there is a highest number x that one has ever added to another number. But how do we know in what way the rule for addition applies to numbers higher than x? Maybe the rule actually was that the addition of a and b equals a+b, as we are used to it, just in case both a and b are lesser or equal to x. But if a or b is larger than x, then maybe the result should rather be a+b+1 or anything else. How can we know has been meant?

to the ones we have encountered already. Therefore, it is possible to decide how we should use our language in these contexts to an important degree. We can then think of a given expression as switching between various related meanings in different contexts. This is done by plurivaluationism of Sud (2020). Nevertheless, this is not enough. The plurivaluationist account reckons only with contexts we know in advance and therewith does not appreciate sufficiently the genuinely dynamic nature of language and rules. The example with dogs and lungs gives us an idea of how the new contexts are open-ended. By the way, this does not mean that all accommodations of a new context are equally good.

Although the idea that something might remain undecided and that meanings are essentially dynamic might seem strange, the idea of an isolated meaning is quite idle and misguided and the first has to be preferred to the second. We will provide a closer description of how this dynamic element in our concepts works but for the moment we see that it cannot be explained away and that it shows how misleading the idea of a completely definite meanings is. This will also modify how we understand ambiguity and the identity of meanings. Ambiguity will become omnipresent, which will mean that a given expression has to be understood as constantly moving in partly unpredictable ways. When asking about the identity of the meaning, this movement will become a part of it. Furthermore, as the identity of meanings will have to recognized as dynamic, the same will have to happen with the identity of contexts. It is a part of the life of language that just as it is not fully clear where a given meaning ends, so it is not clear where a given context stops to apply. But now back to inferentialism, in order to prepare the stage for these ideas.

4. *Caveats in Brandom and how to get off the ground with them – normative attitudes*

Brandom himself acknowledges that meanings cannot be entirely explicit, but it is not clear how he thinks they can work even with that proviso. He brings a useful notion of normative attitude which helps us understand how rules come into life and how they exist thereafter. Normative attitudes are essential to understanding what a rule, and therewith also a meaning, is. Yet we cannot overrate them, as I will try to show.

What is a normative attitude? Primarily it is an attitude a person has towards a kind of behavior. In the most basic form of a normative attitude, the given individual simply considers the given kind of behavior as right or wrong. Thus we typically judge helping the needy as right, as well as drawing inferences according to modus ponens, though right in a different sense. On the other hand, stealing or asserting the consequent are typically deemed wrong in their own ways. Much could be further specified and discussed as to what precisely the normative attitudes are, yet I think one particular point should not be omitted here. Namely, we should not understand normative attitudes as mental states or, at the very least, not as mere mental states which belong to the private sphere of an individual. Assuming a normative attitude should be a public affair, recognizable in one's overt behavior. Such overt behavior can take various guises, yet its basic forms are simply encouraging others to do what we consider right and discouraging them from doing the opposite by sanctions.

Having understood what normative attitudes are, we can examine what their relation to rules and thus also to meaning is. The Brandomian claim is that normative attitudes are constitutive of rules. It is simply by our holding it as such that a rule becomes valid. Just as it can become valid, it can also become invalid. Two basic points have to be emphasized at this juncture.

First, though normative attitudes constitute the rules, these same rules can undergo various developments and these typically cannot be traced back to the specific normative attitudes of given individuals in a society. Thus, the talk about rules certainly cannot be reduced to the talk about normative attitudes. The relation between rules and normative attitudes certainly is not of the straightforward form that we could translate statements about the validity of rules into statements about the normative attitudes in a given society. Yet the two domains are dependent on each other. That is, rules are dependent on normative attitudes. We can make this clearer by making a comparison to Wittgenstein's analysis of the talk of mental states. He dedicates some space to dispelling the notion of a mental state that only the subject can know and which is independent of overt behavior. Surprisingly, though, he admonishes the reader that this all should not be understood as advocacy of behaviorism (see paragraph 307 of Wittgenstein (1953)). Similarly to our case, the talk of mental states cannot be translated into the talk of behavior but it is dependent on it.

After weakening the dependency of rules on normative attitudes, we should also add that the dependency also goes in the opposite direction. Normative attitudes in their more advanced forms depend on rules that are unquestioned in the given context. Any person assumes a normative attitude, besides other reasons, due to her values and the rules she endorses. Yet, despite these caveats, we can say that normative attitudes help illuminate what rules are and how they work.

What do these observations about rules tell us? First, they are not simply here. They have to be kept alive by our normative attitudes. From this follows that they never have a completely definite shape. We have to keep them alive by our attitudes all the time. Though we speak of rules as something that holds, they are rather dynamic entities that have to be resuscitated all the time. It is also in abstraction from normative attitudes that we can petrify them and see them as static. Such an idea probably has its role and is at least a useful fiction, yet in reality, we cannot detach the rules from normative attitudes. This entails both that they are not definite and that they are dynamic.

As far as their dynamic nature is concerned, we can say that the rules are always developing and changing. Even what can be adequately described as *remaining the same* requires our activity and does not come from itself. Every rule enters into new contexts and every application thus contributes something to its content. This does not mean that we cannot in practice, distinguish between establishing the content of a rule and its application, but from a deeper perspective, these two activities cannot really be separated. The content of the rules points to and partly determines their correct application to specific cases. But also, the application to specific cases gives the rules real content. The dependence is mutual.

This dependence of rules on normative attitudes also means that they are bound to remain indefinite, besides being dynamic. Not that we cannot disambiguate, but this can be only partial. In some sense, neither dynamicity implies indefiniteness nor the other way round. Theoretically, one could imagine both the situation that a rule would be dynamic and fully definite and the opposite, namely, a static yet indefinite rule. I, nevertheless, maintain that rules are both dynamic and indefinite.

Let us imagine a rule which would be dynamic yet definite, namely by constantly moving between some specific shapes A and B. Why does this not happen and rules are both dynamic and indefinite? Because new contexts, as I already argued in section 3.3 reveal that some of the applications have not been established yet. Just think of the example with dogs and lungs. Furthermore, there is no way to overview all the normative attitudes, which constantly might push the rule in some direction unthought-of previously. Ultimately, I claim that a completely definite rule does not make sense in a similar, though less obvious, way as the notion of a round square. Of course, some aspects of rules can return, yet in a new context, the return is then imperfect. There is something new added, and therefore we do not fully grasp the shape to which we return. Not all modifications of rules are radical, let alone interesting. But still, we cannot fully fathom where the dynamic will go in advance.

And why cannot a rule be still the same and in addition to that static and still indefinite? As I indicated, the normative attitudes appearing in new contexts just force the rule to move. This is partly because of the necessity to accommodate the attitudes, to bring them into the one flow. Furthermore, although the indeterminacy cannot be fully done away with, we often tend to remedy it. Therefore, the indeterminacy forces specifications and disambiguations that typically give birth to new generations of indeterminacies.

We see that rules are not simply and without further ado determinable and available. They have a very special *modus essendi,* and, as such, cannot be fully identified with any formulation, we can provide. A formulation thus does not truly make the rules explicit as they are because there is no fact of the matter as to their exact shape. Rules, therefore, have a very specific, fluent identity. Questions about where one rules ends and another begins are often meaningless. Or, more cautiously put, one can with equal right say that we have replaced one rule with another, just as we can say that the same rule has developed. And if rules have this specific fluent identity, so have meanings constituted by rules of a specific kind, namely by inference rules.

This lesson about the never vanishing indeterminacy is nicely revealed by Wittgenstein in his musings on rule-following. As there is no determinate way to continue a given number series and as there is nothing to effectively bar deviant interpretations of rules such as Kripke's quaddition, the rules indeed are always in the making. Reading Wittgenstein, we may be unsure whether he speaks only of the complications of getting to know what shape the actual rules have or whether he doubts even the determinacy of the way the rules are. As should be clear, I endorse the second, stronger reading. As it is stronger, asserting it also means asserting the weaker, epistemological interpretation.

Now applying these lessons to meanings, which are constituted by inference rules, we see that they are in the same way indeterminate and dynamic. Of special help is the observation made by Jaroslav Peregrin about what we do when we describe a meaning. When we say that a given word means this and that, it is a special speech act, to use Austinian terminology. It has to capture the actual usage and in this respect it is simply descriptive, yet at the same time, it typically also endorses the very usage. By making such a statement, we encourage the others to use the expression described. This again points to the fact that meaning is never simply here.[6]

Quine and Davidson also hint at the indeterminacy of meaning. Putting aside the differences between these two authors, both radical translation and radical interpretation[7] show us that meaning remains indeterminate. The indeterminacy can be seen as irrelevant in their story of a field linguist wondering how to interpret the unknown word *gavagai,*[8] yet that would be too hasty a conclusion. The radical inter-

[6] Peregrin presents his insight on pages 84 and 85 of Peregrin (2014). The notion of the speech act was, of course, introduced in Austin (1962).

[7] Radical translation is introduced in Quine (1960), while radical interpretation is introduced in Davidson (1973).

[8] The tale of the field linguist and gavagai is also from Quine (1960), chapter two. To remind ourselves, the linguist is trying to understand a language completely unknown to him or her. Furthermore, the language does not resemble any language the linguists has encountered so far. Therefore, she can rely only on the immediate evidence of overt behavior of the community speaking the new language. Now, suppose that the word 'gavagai' is used always in the presence of rabbits. Then it probably means rabbit. Nevertheless, Quine observes that even such a simple case cannot be made conclusively. Maybe the word means rather 'an undetached

preter cannot use other clues besides the overt behavior of community members or, to be more faithful to the scenario, the tribe. The meaning thus observed necessarily oscillates between more shapes, so that *gavagai* can be both a rabbit or merely an undetached part of a rabbit. In this specific context, these differences are immaterial and they do not prevent the interpreter from eventually starting to speak the language and thus enter the linguistic community. Yet this does not mean that the indeterminacy was a mere illusion. It should neither be overrated nor underrated. If we want to hold the meaning fast, it always glides away.

4.1 *Is there a stable core?*

Probably many would agree with seeing language as essentially dynamic, yet would be tempted by the idea that every meaning or at least some meanings must have a stable core that is not subject to change. In the inferentialist framework, this would correspond to the view that although many inferential links between statements can change, some have to remain the same. Peregrin (2014: section 3.6) comes close to this when he distinguishes between *meaning constitutive* inferences and merely accidental inferences. I believe that meanings, in general, do not have stable cores and that we also cannot make the distinction Peregrin does. Let me explain why.

What speaks in favor of reckoning with a stable core of meanings? Meanings must be, to a degree, stable because otherwise, we could not communicate. So much is true, but I believe it is enough if we allow for only relative stability. Some aspects of a given meaning are more central, and therefore the rules which constitute them would be more sorely missed. An example from logic can illustrate this. Although the law of the excluded middle is very important in logic, intuitionists have shown that it is possible to have a logical system that lacks it. Now, one might think that other laws are more fundamental, such as the elimination of conjunction, i.e., the law that states that each conjunct follows from conjunction. Would it even make sense to call an expression *conjunction* if it did not follow this rule of inference? It does not seem probable, but we cannot know all the contexts we will get into. Russell (2018) comes up with counterexamples for the elimination of conjunction. Of course, one is free to doubt the cogency of those counterexamples, but the possibilities to cast even this rule into doubt can hardly be blocked.

Speaking about the meanings of sentences, Recanati (2003: 64), considers whether we need what he calls the *minimal proposition*. His

part of a rabbit' or 'the time slice with the occurrence of rabbits' or something else, related to 'rabbit', yet different from it. It is part of Quine's point that although these alternatives are not the same and there is therefore a genuine dilemma for the linguist, the differences between them might seem as good as irrelevant for most purposes.

background philosophy is different from inferentialism, but he still speaks of a structurally similar problem, namely whether meanings must have immutable cores. And he concludes that it is not clear what positive roles they would play. He gives an example of a mother who tells her child who is crying after a minor injury, ′You are not going to die.′ Taken literally, the mother would be proclaiming the child immortal. But obviously, her utterance is not meant to mean that and this meaning typically would not even occur to the child and would play no role in their exchange.

Furthermore, drawing any such line would not only go against the way language works but would also be very arbitrary. Who is to decide what would belong to the core and what not? Such an arbitrary step should be omitted if it is not necessary. And in this case, I believe we do not need it to secure some language stability. This is because it is enough when the stability is only relative and not absolute. Although I am sympathetic with much of inferentialism, my rejection of the notion of a stable core of meaning and the emphasis on the constant development of meaning make my position differ somewhat from inferentialism of Brandom and Peregrin.

And let us not forget that meaning is inherently holistic according to (by far not only) inferentialism.[9] Meanings are what they are by their relations to other meanings and therewith to the whole of language to which they belong. Then, the same has to be the case for rules that constitute meanings. Therefore, the idea that some inference rules could be given up while others could not be is very problematic. Imagine that a sentence A would obey five inference rules a, b, c, d and e.[10] Now, let us say that, according to the core theory, we can eliminate just the rules d and e, but not a,b and c. But a,b and c are what they are also partly due to their relation with d and e. The rules a, b and c cannot play the role they normally play when they cannot be paired with d and e. And in this role consists what they do and what they are. Therefore, we cannot speak even of keeping the first three rules in such a simple manner. The identity criteria of rules, just like meanings, are very elusive and tricky.

5. *Meaning as constituted only in a particular conversation*

Davidson was led by the phenomena described or very similar ones to a conclusion that might be even more radical than mine. Let us examine his ideas and to what degree we can adopt them. After the period that he dedicated mainly to the idea of radical interpretation, Davidson turned his attention to the specifics of individual conversational situations.

[9] And I believe that this holism is inevitable. Nevertheless, arguing why it is preferable to its possible alternatives, would go beyond the scope of this paper.

[10] From my overall treatment, it should be clear that I do not believe that it is ever possible to exactly enumerate all the rules that a given sentence obeys. This is therefore indeed just a thought experiment.

In his famous *Nice derangement of epitaphs,* Davidson pays attention to the phenomenon of malapropism. We can understand each other despite mistakes we make when speaking or writing. Such a mistake can even be systematic and never corrected by the speaker, yet still, we can manage to understand each other. So much can be readily acknowledged but Davidson seems to draw too strong a conclusion from this observation, namely that there is no such thing as a language:

> I conclude that there is no such thing as a language, not if a language is anything like what many philosophers and linguists have supposed (Davidson 1984: 446).

There are two basic ways of reading this statement. Either we can be led to conceive of communication as something that, completely *ad hoc*, happens only between specific agents here and now, and therefore there is no such thing as meaning shared in a linguistic community. Or we can choose a more careful reading, namely that meaning has a different character than is usually conceived and that the specific situations of the speakers play a much more critical role in the constitution of meaning than one might think.

I choose the second option because there is hardly an explanation of how we can understand someone committing a malapropism besides claiming that the speaker in fact *almost* conforms to the general rules. The guesswork that includes empathy and openness towards others would be hopeless if it could not be embedded into the shared practice and the rules that constitute language. Far from revealing the unimportance of general rules, our capacity to communicate despite malapropisms rather bears witness to their importance. But still, it shows that rules do work in a more sophisticated manner than one could naively suppose.

We must admit that language has to be considered differently, as something less static. A specific situation can indeed bring a lot about how we understand each other. It would be tempting to say that what is specific for a given situation does not concern meaning itself but rather how we manage to grasp it in a given context. Such a remark is not entirely illegitimate but there hardly can be a firm line between what indeed belongs to language and what only belongs to a given situation. Taking such a boundary all too seriously would ignore the lessons we learned from Quine (1951) and others a long time ago about the untenability of the firm distinction between analytic and synthetic statements and truths. Indeed, this strategy of explaining Davidson's insights away would amount to claiming that what is specific about understanding each other here and now is that we have to gain the synthetic knowledge, the lack of which might prevent us from grasping what the other person means. It would also come close to endorsing some inference rules as meaning constitutive, which is a position I have already argued against. Every inference rule is, to some degree, meaning constitutive. But every rule is also revisable. When we communicate, we understand that language is dynamic and relative stability is enough for us.

Here I again agree with Recanati (2003), who endorses contextualism as an alternative to literalism. In any real conversation, we have to heed to its specifics and thus pragmatically modulate our understanding of what is said. Meaning is thus created during the conversation. Of course, we enter every conversation somehow prepared and have some idea of which rules hold and what specific expressions typically mean. But no cogent boundary can be drawn between the pragmatic modulation and what it modulates.

The refusal of the stable core also means that the contexts are not wholly available to us in advance; we cannot evaluate the specific situation as forming a context for which we have antecedent rules. Although what we know already is an essential guide in the new context, this context is radically new and its rules have to be formed yet. In this way, the notion of ambiguity as the list of possible meanings of a given expression is not fortunate. In advance, we have a tentative list of possible meanings that are to be revised continually. And every expression is, in this sense, ambiguous that its meaning can modify. The general notion of ambiguity thus fails to delimit a specific set of expressions.

5.1 *What one means*

Davidson (1984) frames his account in terms of how we manage to guess what our interlocutor means. It would be futile to refute the obvious, namely that when speaking with someone, we try to find out what our interlocutor has in mind. It is quite common for us to ask the others what they mean and treat them as those who bear the meaning hidden inside them and are trying to convey it to us.

But caution is necessary from the very beginning of this debate. In fact, putting too much weight on what is inside a given individual renders the debates about meaning impossible. Only if we consider what the interlocutor means as somehow accessible to the others, can we rescue the intelligibility and rationality of the account. But furthermore, even if somebody can be in a specific state of mind when uttering a sentence, this does not mean that the state of mind determines what the sentence means. Not only what it means in general but even what it means in his or her mouth, in the specific situation. It is quite common that we only subsequently discover what we actually expressed when we said this and that. The meaning continues being created and formed in the course of the conversation. It would be misguided to think of it as something that was ready from the very beginning and only had to be transmitted.

The specific state of mind of the speaker certainly plays a role in a specific conversation, yet it does not determine what the specific utterances of the speaker mean. My account is thus rather far away from psychologism. Having something on one's mind and intending to say something is only a preparation to begin the linguistic interaction and therewith also to let any meaning come into play. From a certain point of view, this is quite an obvious observation, yet identifying meaning with what one merely means contradicts this very platitude.

5.2 *Modifying meaning*

The Davidsonian insight that meaning is constituted only in a specific conversation thus has to be taken with great reserve. Yet it can help us start describing the mechanism of how meaning changes and how we can start such a change. This was undertaken by Ludlow (2014) and a very good summary can be read in Drobňák (2017). In order to initiate a change of meaning, it does not suffice to start using a given expression differently than people use it normally. You also must have a certain authority in the community and the modification you are pushing forward has to be reasonable in some ways. Typically, you should be able to provide arguments in favor of this change.

My account emphasizes the dynamic side of meaning but its stability must also be appreciated. Changing meaning is no simple affair and cannot be done by simple fiat. By far, not every conversation that people have leads to groundbreaking changes in how we use our language. Lots of changes instead need an inspiring charismatic personality to realize them. It can be a politician who starts using an old expression in an inspiring new way or coins an altogether new expression; it can be a popular singer or a genius author such as Shakespeare. Some of the changes might be seen as fortunate, others can be harmful but we will not examine how these can be distinguished. At any rate, what any such modification brings can hardly be foretold before it actually begins to function and live in the community.

6. *Elusive meaning*

Although it is very important that we can decide to change the meaning of an expression under appropriate conditions, I focus rather on a different phenomenon. Namely, that meaning that can never be fully explicit is never fully in our hands. It is always unstable and tends to change automatically as we use it. Such a spontaneous change can take a long time to happen, lots of individual dialogues it is featured in bring gradual and imperceptible modifications until we realize that a qualitative change has occurred. At least, when we succeed in making this change explicit.

The fact that meanings are never fully explicit and thus never fully in our hands should not be treated as a passing observation but as a fundamental feature. Only in this way can language live with us in the constantly changing world. This also differentiates my view of language and meaning from Ludlow (2014) and Cappelen (2020). Both these authors share my general attitude regarding language as dynamic. Nevertheless, from my point of view, they regard language as too much in our control. I, in contrast, believe and will yet illustrate in the next section, that we in many ways do not decide what is correct and what is not and that, therefore, language is in an important way independent of us. Ludlow speaks a lot of how we negotiate the meanings of expressions in a given conversation, while Cappelen in a related manner speaks of conceptual engineering. This means that we can improve our concepts by attaching pragmatically better meanings with our expressions. I do not doubt that we negotiate about what we mean in a given situation, as I witnessed by my sympathies for Davidson who went in a similar direction. I also do not doubt that we can try to change the meanings of our expressions and have good reasons for it. But nevertheless, meaning is, in my view, dynamic even if we do not actively try to change it.

I shall illustrate this difference in the next section. I believe it can be instructive to understand language and individual expressions as living beings. A closer illustration of what this can mean is provided in the philosophy of Henri Bergson.

6.1 *Bergsonian meanings*

Before presenting how Henri Bergson can help us understand rules and meanings better, we will have to review some of the basic tenets of his overall philosophy, though very briefly. Bergson considers reality to be fundamentally dynamic and living. He considers movement as the veritable foundation of all we experience, although we tend to overlook it and consider it as secondary in an essentially stable world.[11]

According to Bergson, movement is a fundamental happening in the world and is irreducible. When observing a given object moving from point A to point B, we can describe the positions between them which it successively goes through but the movement itself is something over and above these mere positions, indeed it is even something completely different. Reducing it to these positions amounts to banishing movement from the picture altogether. When we consider one of these positions occupied or one of the stages of the movement, we in fact abstract the movement from our consideration and render the moving object stable. The temptation to do so is, among other things, an important source of Zeno's paradoxes. Concerning the arrow paradox, we indeed conceive the flying arrow as motionless when we consider it in the individual positions. By the same token, Achilles can never surpass the turtle in the race. At least when we think of the positions occupied during the movement or stages of movement rather than of the movement itself. That is, when we falsely try to reduce movement to its stages which are static phenomena.

[11] I am only giving a sketch of these basic Bergsonian motives. A good introduction to them is Bergson's first major work, Bergson (1889).

A dynamic phenomenon that Bergson pays particular attention to is consciousness. We tend to see it as a succession of certain states, be they emotional or intellectual. Though, such a conception is a distortion, as consciousness is essentially a flow. Even what seems to be remaining in the same state is a kind of movement.[12]

We tend to take a misguided perspective on dynamic and living phenomena because we misconceive time. Science understands time as a further dimension of space and thus neutralizes it. Space is the locus of homogeneity and stability, while real, original time, which Bergson calls *durée*, is characterized by heterogeneity and a dynamic structure. It is unpredictable what will happen, and the new is bound to be radically new.

In this way, I believe, we should also consider language to be a living phenomenon. But why does it happen that we tend to disrobe reality of its dynamics? According to Bergson, it is no simple mistake, from a certain point of view it is a reasonable way of coping with reality. In this way we consider reality to be something predictable and stable and thus can much better plan our actions and predict their outcome. Such a utilitarian perspective is indispensable but should at the same time not be considered as absolute and to be the only one.

As it can render our lives more agreeable, it can also flatten and make us forget who we are and what language itself is. In the case of meaning, which is our topic, it causes us to see it as a ready, static and dead thing. But we have seen arguments as to why meaning cannot be fully explicit and thus in our control and change only if we decide to change it.

This by no means amounts to advocacy of linguistic anarchy. Clearly, there are rules and I believe inferentialism reveals much about the meaning and what it is. Yet those rules are much more fluid than they might seem and this should not be ignored or abstracted from. And the line between breaking these rules and inventing new ways in which our language can function is also very unclear and unstable. Enriching inferentialism by this Bergsonian element leads to what might be called *dynamic inferentialism*.

Bergson (1932) speaks of two ways our rules can force themselves on us, one being coercion, the other aspiration and *élan*.[13] The first force

[12] Bergson's account of consciousness thus radically differs from Hume's bundle theory of the self from Hume (1738). Bergson's position is succinctly formulated in this quote Bergson (1907: 11): "Il est commode de ne pas faire attention à ce changement ininterrompu, et de ne le remarquer que lorsqu'il devient assez gros pour imprimer au corps une nouvelle attitude, à l'attention une direction nouvelle. A ce moment précis on trouve qu'on a changé d'état. La vérité est qu'on change sans cesse, et que l'état lui même est déjà du changement." [Translation, p. 2: But it is expedient to disregard this uninterrupted change and to notice it only when it becomes sufficient to impress a new attitude on the body, a new direction on the attention. Then, and then only, we find that our state has changed. The truth is that we change without ceasing and that the state itself is nothing but change.]

[13] Bergson (1932: 53): "Dans la prèmiere, l'obligation represente la pression que les éléments de la société exercent les uns sur les autres pour maintenir la forme

mainly serves to preserve and conserve the rules we have, the other serves to give them new life. We need both these and both should be reckoned with if we want to understand how language works. Therefore, the identity of a meaning of a given expression cannot consist merely of a set of rules, although such a set is an important ingredient. The other essential ingredient is the irreducible movement. This movement does not break the identity of a given meaning, it rather necessarily belongs to it. By movement and change, we are not obliged to speak of a different meaning. The irreducibility of movement to its stages, no matter how fine grained such a reduction would be, entails that meaning of a living expressions cannot be reduced to many precise meanings. This is because the precise meaning is an illusion caused by forgetting the dynamism which is always present. My account thus differs, as I already noted, from plurivaluationism of Sud (2020) which claims that we typically speak many precise languages at the same time. Such a perspective can be useful but we also need the opposite perspective, namely that we typically speak one, though dynamic and living language.

7. *Making meanings explicit and caring about language*

Besides inferentialism, Brandom is also a proponent of the related idea of logical expressivism. As we already know, inferentialism considers inference rules as constitutive of meaning. At the same time, Brandom himself admits that these rules are often not explicit. I agree, though I argued that in a strict sense no meaning is ever fully explicit, indeed the very idea of a fully explicit meaning is misguided.

Logical expressivism claims that logic with its vocabulary is here to make the inference rules explicit. Let us come back to our example of inferring that Rex is a dog from his being a dachshund and further inferring that he is a mammal. These inferences are correct due to the rules stating that every dachshund is a dog and that every dog is a mammal. These rules are rendered explicit due to the logical vocabulary, such as the word *every*.

I believe that when Brandom considered logic as a tool for making inference rules and therewith meanings explicit, he was on the track of an important idea. But it should be added that the meaning cannot just be rendered explicit as it was because we modify it slightly by rendering it explicit. It is valuable to express the inference rules that regulate a given expression and try to express them continuously with the

du tout... Dans la seconde, il y a encore obligation, si l'on veut, mais l'obligation est la force d'une aspiration ou d'un élan, de l'élan même qui a abouti à l'espèce humaine." [Translation: In the former, obligation stands for the pressure exerted by the elements of the society on one another in order to maintain the shape of the whole... In the second, there is still obligation, if you will, but that obligation is the force of an aspiration or an impetus, of the very impetus which culminated in human species.]

actual usage. But any such expression at least stabilizes the meaning, impedes its natural movement. That is already a change of the meaning as it was before. Therefore, there is nothing as pure expression, free of any modification of what it expresses.

Indeed, logic with its vocabulary and language in general is, among other things, a force of stabilization, even rigidization. This is the light in which Bergson typically characterizes both language and logic. As institutions that not only describe the world as stable but even render it such. I am proposing to extend the Bergsonian appreciation of the dynamics of the world we live in even to language and consider it a dynamic, living entity constituted by the normative attitudes of a given community.

When we make a rule explicit, we act as if this rule was a ready and firm part of the meaning we just brought to the fore. Yet we also help to render it valid. Making rules and therewith meanings explicit is thus not just theoretical observation of the meanings and how they are but rather a specific form of interaction with them. Such an interaction is an important part of our freedom concerning language and it is correct to take advantage of that.

On the other hand, it should be acknowledged that clear-cut meanings are chimerical and that meaning is always dynamic and to some degree elusive. The search for full clarity is thus misguided and we should treat language with due respect, which also means acknowledging that it always partially escapes our control. That is the reason why it is such a fascinating thing and it would be a great pity if it were otherwise.

Let me note that classical inferentialism of Brandom or Peregrin does leave some space open for the dynamic character of meaning but it is not enough in my view. Indeed, Brandom acknowledges that meaning is partly *perspectival*.[14] Peregrin (2014: 51) notes that when someone says, "The man over there left the room with blood on his hands," then clearly, someone who believes that the person is a doctor who has just finished an operation understands this sentence differently than someone who thinks that a murder is being described. But on this view, we still have to choose from a completely firm and stable basis of inference rules and apply those that suit the given context. When a given sentence is paired with one set of premises, it enables us to judge something different than when it is paired with a different set. But this is just an illustration of the point that meanings are constituted by inference relations rather than somehow intrinsic to a given expression. This point might be seen as a good first step of inferentialism towards understanding meanings as live and dynamic but a much longer path has to be undertaken to make inferentialism indeed appreciate the true dynamics of meaning. On the view presented here, it is not only the choice of relevant inferential relations that has to be taken into account but rather the fact that these very relations change

[14] See Brandom (1994: 594–597), where he discusses how meanings can be both perspectival and objective.

their character and develop. As I already said, the notion of inferential potential, IP, is a good model but should not be taken all too literally. IP only gives us an idea of what gets changed and modified all the time when language is used.

Conclusion

We started by considering the identity of an expression or its meaning. We did this in particular using the idea of an expression that is ambiguous concerning its meaning. This idea is naturally paired with its opposite, that is with the idea of an expression to which just one clear meaning is associated. Meaning and its identity are more complicated than expected, as should be abundantly clear from the course of my considerations. Rather than being associated with a specific set of rules or with a specific shape, the meaning is constituted by its history. Rather than being a thing, it is a happening, a process. And as such, it hardly possesses any clear criteria of identity. Maybe one could speak of dynamic criteria of identity for dynamic expressions, in line with *dynamic inferentialism*. The dynamic and indeterminate criteria of identity of meanings is mirrored by the criteria of identity of contexts, which possess the same characteristics.

Understanding an expression thus amounts not so much to readiness to give a satisfactory definition, although it can be manifested by such a readiness, but rather by the ability to participate in the very happening which it is and its history. It also amounts to taking a certain responsibility for how we use the expression and develop it in the new contexts that both we and the expression in question enter into. It should also be clear that if the notion of definite meaning has to go, then so does that of ambiguity, which is just its reversed side. Or at the least, it needs to be rethought anew.

It makes perfect sense to characterize a given expression as ambiguous, as contrasted to the tidier ones in specific cases and contexts. Sometimes, it is also meaningful to consider such expressions problematic. But this is a perspective of a more practical linguist. From a philosophical point of view, ambiguity pertains to all expressions, though in variegated ways and degrees. From my point of view, the notion of ambiguity is therefore not very useful, as it is a feature of all expressions and thus does not delimit an interesting class. However, it points to the necessity of regarding every meaning as a dynamic and living entity.

References

Austin, J. L. 1962. *How to Do Things with Words*. Oxford: Oxford University Press.

Bergson, H. 1889. *Essai sur les données immédiates de la conscience*. Paris: Felix Alcan. English translation by F. L. Pogson (1910). *Time and Free Will: An Essay on the Immediate Data of Consciousness*. Montana: Kessinger Publishing Company.

____ H. 1907. *L'Évolution créatrice*. Paris: Felix Alcan. English translation by A. Mitchell. 1922. *Creative Evolution*. London: Macmillan.

____ 1932. *Les Deux Sources de la Morale et de la Religion*. Paris: Felix Alcan. English translation A. Audra and C. Brereton: *The Two Sources of Morality and Religion*; University of Notre Dame Press, 1977.

Brandom, R. 1994. *Making it Explicit*. Cambridge: Harvard University Press.

Cappelen, H. 2020. "Conceptual engineering: The master argument." In A. Burgess, H. Cappelen and D. Plunkett (eds.). *Conceptual engineering and conceptual ethics*. Oxford: Oxford University Press, 132–152.

Carnap, R. 1928. *Der Logische Aufbau der Welt*. Berlin: Weltkreis. Translated into English as *The Logical Structure of the World*. Berkeley: University of California Press.

Davidson, D. 1973. "Radical interpretation." *Dialectica* 27 (1): 314–338.

____ 1986. "A nice derangement of epitaphs." In E. Lepore (ed.). *Truth and interpretation: Perspectives on the philosophy of Donald Davidson*. Oxford: Blackwell.

Drobňák, M. 2017. "Meaning-constitutive inferences." *Organon F* 24 (1): 85–104.

Hume, D. 1738. *A Treatise of Human Nature*. New York: Oxford University Press, 2000.

Keefe, R. 2000. *Theories of Vagueness*. Cambridge: Cambridge University Press.

Kripke, S. 1982. *Wittgenstein on rules and private language: An elementary exposition*. Cambridge: Harvard University Press.

Lepore, E. 2007. "Brandom beleaguered." *Philosophy and Phenomenological Research* 74 (3): 677–691.

Ludlow, P. 2014. *Living words*. Oxford: Oxford University Press.

Peregrin, J. 2014. *Inferentialism: Why Rules Matter*. London: Palgrave Macmillan.

Quine, W. v. O. 1951. "Two dogmas of empiricism." *Philosophical Review* 60 (1): 20–43.

____ 1960. *Word and Object*. Cambridge: MIT Press.

Recanati, F. 2003. *Literal meaning*. Cambridge: Cambridge University Press.

Russell, G. 2018. "Logical nihilism: Could there be no logic?" *Philosophical Issues* 28 (1): 308–324.

Sellars, W. 1974. "Meaning as functional classification." *Synthese* 27 (3): 417–437.

Smith, N. J. J. 2008. *Vagueness and Degrees of Truth*. New York: Oxford University Press.

Sud, R. 2020. "Plurivaluationism, supersententialism and the problem of many languages." *Synthese* 197 (4): 1697–1723.

Williamson, T. 1994. *Vagueness*. London: Routledge.

Wittgenstein, L. 1953. *Philosophische Untersuchungen*. Oxford: Blackwell.

Croatian Journal of Philosophy
Vol. XXII, No. 64, 2022
https://doi.org/10.52685/cjp.22.64.5
Received: December 20, 2021
Accepted: May 10, 2022

Empty Higher Order States in Higher Order Theories of Consciousness

SINEM ELKATIP HATIPOGLU
Marmara University, Istanbul, Turkey

According to higher order (HO) theories of consciousness, a mental state is conscious when there is a HO state about it. However, some HO states do not seem to be about other existing mental states. It is possible to resolve this problem since targetless HO states resemble HO states that misrepresent but the assumption that HO states always target other existing mental states is at odds with the theory since HO states are not only necessary but also sufficient for phenomenal consciousness according to the theory. Given the sufficiency of the HO states for consciousness, there is a need to understand the emergence of HO states as a non-random phenomenon to avoid the difficulties caused by targetless HO states. I suggest it is possible to develop such an understanding by thinking of HO states as predictive states in accordance with the predictive processing theory of the mind.

Keywords: Consciousness; higher order theories; empty higher order states; predictive processing.

Introduction

According to higher order (HO) theories of consciousness, a mental state is conscious when its subject is aware of it in a suitable manner. Among different accounts of this awareness (see Gennaro 2004 for an overview), one view states that the awareness involves a mental state that is distinct from the mental state one gets to be aware of. For instance, according to Rosenthal's (2005) higher order thought (HOT) theory, a mental state is conscious when its subject is aware of the state by way of having thought about it. The mental state one gets to be aware of is the target state or the lower order (LO) state and the thought about it is the HO state.

It is important to note that according to the HOT theory, HO states are not only necessary but also sufficient for phenomenal consciousness. I will refer to this as the *sufficiency principle.* Accordingly, the subject does not necessarily have to be in some LO state for the HO state to represent its subject to be in that LO state. Also, even if the subject is in some mental state, it does not necessarily follow that she will be phenomenally conscious of it in the absence of a HO state. This "division of phenomenal labor" between the LO and the HO state has been a source of criticism directed at HO theories.[1]

Higher order states may accurately represent the mental state that the subject is in, or misrepresent it, or represent the subject to be in some mental state that she is not even in. Criticism of the division of phenomenal labor is particularly powerful in this last case, viz., the case of empty HO states where there is a HO state without a target state (see for instance the discussion between Block 2011a, 2011b, Rosenthal 2011 and Weisberg 2011a and 2011b). I refer to this criticism as *the empty HO state objection.* In the case of empty HO states, there is an additional concern about which particular state is conscious in virtue of the HOT since the possibility of an empty HO state shows that the theory is committed to saying that subjects can be phenomenally conscious of mental states that they are not in.

Wilberg (2010) emphasizes this particular problem when he raises the question of which existing token mental state is conscious in virtue of the empty HOT and finds Rosenthal's suggestion that the conscious mental state "... may be a merely notional state and may not actually exist" (2000: 232) to be in conflict with the fact that his theory is a theory of *state* consciousness, according to which consciousness would be a property of a mental state token. Wilberg denies that consciousness is only a matter of appearance and consequently denies that when it seems a certain way to a subject then she must be in a conscious state. Otherwise, one would be forced to simultaneously say that the mental state token exists and does not exist in the case of empty HO states. The mental state token does not exist because the HOT is empty and it exists since it seems a certain way to the subject, i.e. it seems to the subject as if she is in a specific mental state. To remove this incoherence, Wilberg's account of empty HO states consists in what he calls the "no consciousness account" according to which a subject is not in a conscious state in the case of empty HO states.

Berger (2014) undermines Wilberg's (2010) argument for incoherence and reinstates the notion of consciousness as a matter of appearance, more specifically as a matter of which mental state it seems to oneself to be in. Hence according to Berger, one's awareness of a mental state strictly speaking should be understood as one's awareness of

[1] This critique of higher order theories was first taken up by Byrne (1997) and then by Naender (1998), and Levine (2001). Later, others such as Kriegel (2003) and Mandik (2009) have addressed the same issue. The phrase "division of phenomenal labor" appears in Naender (1998).

oneself as being in a mental state. As such, he argues that despite the terminology of state consciousness, the property of consciousness really attaches itself to individuals (2014: 831). Therefore, Berger says that there is no problem with empty HO states if consciousness is taken as a property of subjects and not existing mental states.

Block (2011a) makes a distinction between the ambitious and the modest version of HO theories and contends that when faced with the question of why putting together an unconscious pain with an unconscious thought about it results in a conscious pain, the ambitious theory must provide a meaningful answer since unlike the modest view it aims at an account of the nature of what it is likeness. He then argues that the HOT theory cannot achieve this because it abuses the notion of what it is likeness as can be seen in its response to the empty HO state problem. Block (2011a: 426) says that "If what it is likeness is supposed to *matter* in the same way *whether it exists or not*, that just shows that 'what it is like' is being used in a misleading way" (his italics).

Farrell (2017) argues that if empty higher order states are endorsed by HO theories, then one should deny that these theories account for what-it-is-likeness, and without such an account a theory of consciousness is no longer an ambitious theory. According to Farrell, to undermine the problem of empty HO states and of misrepresentation for that matter, HO theorists adopt what he calls an occurrent reading of there being something it is like for the subject to be in a mental state (2017: 2748) and a loose reading of there being an occurrence of what-it-is-likeness associated with a mental state (2017: 2750). According to these readings, there being something it is like for the subject to be in a mental state entails that there is an occurrence of what it is likeness associated with that mental state but the subject does not have to be in some mental state for there to be a what it is likeness associated with that mental state. Farrell then argues that neither of these readings fit with our ordinary conception of consciousness based on the Nagelian definition and therefore HO theorists either would not really be responding to their opponents' arguments in adopting these readings or they become non-ambitious theories of consciousness since they cannot provide an account of what-it-is-likeness.

Gennaro (2012) tries to resolve the issue by developing another version of the HOT theory, viz., WIV (Wide Intrinsicality View) theory, according to which the HO state is actually a part of the lower order state and together they form a complex conscious state, hence the LO state is not numerically distinct from the HO state.

I contend that the empty HO state objection arises as a consequence of not taking the sufficiency principle seriously enough and relies on the false assumption that a HO state must target a LO state. As Rosenthal (2000: 232) points out, the so-called LO state can be a non-existent or a notional state. However, the sufficiency principle is not welcomed because there is not enough literature discussing the emergence of HO states or how they may be related to LO states when and if they are related to them.[2] Thus, the emergence of the HO states seems like a random phenomenon that further fuels the empty HO state objection. By providing some theory about the emergence of HO states, both the dichotomy between the LO and the HO state and the sufficiency principle would be better understood. While the HO theory that I focus on is Rosenthal's HOT theory, most of the things discussed here are relevant to any HO theory where the HO state is distinct from the LO one.

In this paper, I first discuss a way to undermine the empty HO state objection which relies on the arbitrariness between the empty HO state phenomenon and misrepresentation and then explain what is wrong with this approach. The right approach should be compatible with the tenets of the theory viz., the sufficiency principle. This, I suggest, is possible by taking HO states to be similar to predictive states in accordance with predictive processing theory of the mind (see Clark 2016, Metzinger and Wiese 2017). My purpose is not to develop a complete theory of the HO states as predictive states but only to pave the way for a theory of the emergence of HO states.

1. *The so-called accurately represented targets, misrepresented targets and absent targets*

Consider the following examples according to which I am in a,

(1) LO mental state of seeing a green apple
(2) LO mental state of seeing a green ball
(3) LO mental state of seeing a red bowl

According to the HOT theory, it is possible for the subject to have a HOT with the content "I'm seeing a green apple" in all these cases and be phenomenally conscious of seeing a green apple. I follow Weisberg (2011a: 416) in calling a case like (1) veridical representation, (2) misrepresentation and (3) as involving a HO state with no target or an empty HO state.

Wilberg (2010) says that it is possible to understand cases of misrepresentation as cases where the target of the HOT does not exist. Similarly, Rosenthal (2004: 32) finds the "distinction between an absent target and a misrepresented target … arbitrary" and says,

> Suppose my higher-order awareness is of a state with property P, but the target isn't P, but rather Q. We could say that the higher-order awareness misrepresents the target, but we could equally well say that it's an awareness of a state that doesn't occur. The more dramatic the misrepresentation, the greater the temptation to say the target is absent; but it's plainly open in any such case to say either.

[2] When it comes to the relation between HO states and LO states, Rosenthal (1993a) denies it to be causal and the best scenario is that of an accompaniment. This being the case, one wonders if there is any limit—and on what grounds to the way a HO may represent a LO state.

One may then say that if misrepresentations are unobjectionable, so should empty higher order states be. However, I contend that this is not the right approach to defend the HO theory from empty HO state objection. While it may be tempting to resolve the issue about which token state gets to be conscious in the case of empty HO states by likening absent targets to misrepresented ones and thereby assigning targets to them, such an approach only reinstates the assumption that a HO state always targets a LO state. This would overlook the sufficiency principle.[3]

The mental state the HO state represents its subject to be in may coincide with a certain existing LO state that the subject is in but this is neither a necessary aspect of the theory nor is it a necessary feature of the relation between the HO state and the LO state, assuming there is a relation. It is interesting to note that the so called veridical cases where the notion of 'HO state targeting a LO state' is perhaps the most powerful may also be redescribed in a way to involve misrepresentations. For instance, in the case of (1) the subject may be phenomenally conscious of seeing the apple's color as a generic green rather than the particular shade of green the LO state represents the apple to have. The orthodox way of thinking about this is usually as a case of veridical representation where certain subtleties are lost in the HO representation of the LO state. However, given the sufficiency principle, it is just as reasonable to think of the HO state independently of the LO state. Similar to Rosenthal's earlier suggestion one might say that the subject's HO awareness is of a state with property P (generic green) but that the target isn't P but rather Q (the particular shade of green). Hence one might suggest that there is a certain sense of arbitrariness concerning the distinction between veridically represented LO states, misrepresented LO states and absent LO states. While that may be true, it would be wrong to use this idea and contend that there are no empty HO states to undermine the empty HO state objection since the idea relies on the false assumption that a HO state must always target an existing LO state.

2. *Randomness and empty higher order states*

Resembling the empty HO state phenomenon to misrepresentation and thereby rendering the HO state non-empty reinstates the idea that a HO state must always target a LO state and therefore is at odds with the sufficiency principle. If the sufficiency principle is dispensable for HO theories, then the above approach might work but I don't think it is dispensable. Hence the HO theorist needs to address why the suf-

[3] I'm not suggesting that Rosenthal's purpose (2004: 32) in the quotation above is to undermine the empty HO state objection based on the arbitrariness. Instead it should be understood as providing some clarification on the notion of an absent target or an awareness of a state that doesn't occur.

ficiency principle which is really at the core of the empty HO state objection is not welcomed.

As Gennaro (2012: 60) and before him Levine (2001: 108) have discussed,[4] since HO states are sufficient for phenomenal consciousness, there seems to be no point of there being a lower order state, especially a numerically distinct one.[5] The presence of an actually existing LO state is possible but neither necessary nor sufficient for consciousness and empty HO states stand out because they make the sufficiency of HO states for consciousness more obvious.

Without articulating why and how HO states come about, an inevitable sense of randomness threatens the theory. The concept of an absent target makes this randomness obvious, while a misrepresented target promises a story about how the HO state is still about the LO state and fits better with our general understanding of mental lives by making them seem less random. It seems that this sense of randomness fuels the empty HO state objection. Therefore, one could be tempted to argue that there are no genuinely empty HO states but only misrepresented targets. However, as mentioned before, I consider this at odds with the very tenets of the theory.[6]

If randomness is to be avoided and some theoretical background is to be provided for the emergence of HO states, I'd like to suggest that this is possible by incorporating the mental history of the subject into the emergence of HO states. While LO states may be a part of that history, they would not stand out in any special way in terms of their relation to the HO states.

I will not try to articulate in detail what the mental history of a subject refers to but it is meant to be the kind of thing that gives rise to the phenomenal differences between for instance Mary's[7] first experience of seeing a red chair after leaving the black and white room and her

[4] Gennaro (2012: 60) says that one faces the question of "… what the point of having both a LO and HO state is if only one of them determines the conscious experience." Likewise, Levine (2001: 108) says that "the first-order state plays no genuine role in determining the qualitative character of experience."

[5] This is probably why Gennaro (2012) develops his version of a HO theory of consciousness, viz. WIV (Wide Intrinsicality View) theory according to which the higher order state is a part of the lower order state and together they form a complex conscious state, hence the LO state is not numerically distinct from the HO state. My purpose is to assess the empty HO state objection for theories where the LO state is numerically distinct from the HO state, not when it is a part of the HO state. Obviously, the question whether WIV is able to tackle the empty HO state objection while remaining to be a higher order theory is worth examining but I cannot undertake this task here.

[6] Besides, even if absent targets are replaced by misrepresented ones, one still faces the question of why HO states would misrepresent their targets in this radical way or in what sense a HOT with the content 'I'm seeing a green apple' would still be about 'perception of red bowl' LO state.

[7] Jackson's (1986) example of the super scientist who is omniscient concerning physical knowledge and knows all about colors but was grown up in a black and white room and has never seen a colored object before.

experience of the same red chair two years after she leaves the room. These differences, while taken for granted, are not addressed sufficiently with the exception of Rosenthal (1991: 33-4, 2002: 413-4) who argues that one of the advantages of the HOT theory is in its ability to explain how one's conceptual resources influence the phenomenological features of one's experiences since the HO state is a thought.[8]

Consider the following example of the impact one's mental history has on one's consciousness. A woman is sitting in the lobby of a building, waiting anxiously to meet her long-lost brother. She is constantly checking the sliding doors that open to the lobby. Then a strong wind causes a plastic bag to fly in. It is conceivable that being phenomenally conscious of seeing her brother, the woman gets off her seat to meet him and soon realizes it was just a plastic bag. Perhaps one might suggest that given the 'perception of the plastic bag flying in' as the LO state, along with the desire to see the brother, the anticipation etc., the brain is in some sense forced to predict that her brother has arrived resulting in the HOT 'I'm in a mental state of seeing my brother.' One might even suggest that this prediction is for the organism's well-being, for instance, to momentarily reduce the stress the subject suffers from. Hence even though the HO state is targetless, there is a certain background, a certain mental space in which this particular HO state comes about.

This is a direction in which HO theories may further be developed, viz., by providing an account according to which empty HO states arise for the organism's well-being given its mental history. This would eliminate randomness for two reasons. Firstly, empty HO states would be driven by a purpose. This purpose could be to sustain a certain level of equilibrium in the subject's mental life by avoiding too much stress. It could be a reaction to the mental history of the subject. Just as blinking is a physical reaction to protect one's eye when something gets close to it, empty HO states could be a mental reaction to protect one's mental health under conditions where the subject needs to have an experience *x* even though she is not in that particular mental state *x*. Secondly, empty HO states would be grounded in some mental space rather than being randomly generated since they arise in relation to the subject's history.

In fact, it is possible to think this way about HO states in general and not just empty ones. One way to do this is to think of HO states as the predictive states in accordance with the predictive processing theory of the mind (PPT) (see Clark 2016, Hohwy 2013, Metzinger and Wiese 2017). Given the subject's history, the HO state's representation of its subject to be in some mental state would actually be a prediction of what the subject would be phenomenally conscious of. PPT emphasizes the constructive nature of mental episodes, such as perception and the top-down processing that is involved. Hence perception is not merely

[8] He gives the example of wine tasting, musical experience (1991: 33–4) and the experience of hearing the sound of an oboe (2002: 413–4).

passive and stimulus-driven. Instead, it is active and also hierarchical. This top-down processing is not something that is effective only when sensory input cannot be relied on but it is essential to and constructive of perception. Put simply, the brain makes use of computational models in accordance with Bayesian inference as a computational method to make predictions about the external world that the subject is in and the possible causes of the effects that the subject is receiving information about through sensory signals. A more dramatic way to put this is to say that the brain dreams in a world where dreaming is not random but very much controlled (Metzinger 2003: 52).

The next step, again put simply, involves the brain asking to itself if the prediction it's made is correct. This is done by taking into account the sensory input and checking if there is a mismatch between the sensory input and the prediction, provided that the sensory input is reliable. If the sensory input is not to be trusted, i.e., if it is too noisy or ambiguous, even if there is a mismatch, the sensory input is undermined and prevented from being further processed. However, if the sensory input is reliable and there is a mismatch between it and the prediction, the computational model that the brain uses to make its predictions is revised to decrease errors in future predictions.

Given the sufficiency principle and the relevance of conceptual resources to one's phenomenal consciousness in HOT theory, granted that the HO state is a thought, I contend that the HOT theory of consciousness is the most compatible one with PPT since it allows for the top-down process that PPT endorses rather than a bottom-up process. Interestingly enough, the evidence for this lies in the phenomenon of empty HO states even though empty HO states are usually the source of an objection to HO theories, as discussed in the beginning. The simple fact that being in a conscious state does not necessarily involve being in that state in the HOT theory may be seen as evidence for the top-down process. Just as the sensory input in perception according to predictive processing is used to check if the prediction is correct, and therefore not initially essential to the prediction in the top-down framework, the mental state that the subject is in may be considered to be non-essential to the HOT about it but may later be used to check the accuracy of the HOT, that is if the subject is indeed in such a mental state.

So instead of the subject being in a mental state and there being a HO representation of that mental state the subject is allegedly in, which would be a bottom-up process, the HO representation can be taken to be a prediction of the mental state the subject would be in regardless of whether or not the subject is in that state. Hence technically, the HO representation would not be the representation of a LO state strictly speaking but a thought of a predictive nature about some LO mental state that the subject might be in, given the circumstances. And again in accordance with PPT, the next step would involve checking if the prediction is correct. In the example given above, since the subject does not see her brother in the moments that follow as she approaches the doors, the prediction would need to be corrected. If empty HO states are typically taken to be rare or not to last long, this prediction-checking followed by a revision when needed would provide an explanation for the rarity or short duration of empty HO states. Another analogy that may be observed between PPT and HOT theories is that just as the predictions are not experienced as predictions by the subject, the HOTs are not typically conceived as thoughts that the subjects are conscious of having.[9]

This way of thinking about the emergence of empty HO states, or HO states in general calls for a change in our ordinary ways of thinking about consciousness which usually include a bottom up process of being in a mental state and then being aware of being in it. However, as Rosenthal (2004: 41) notes, consciousness is not actually about being in a state and being conscious of being in it. The first part of this conjunction is in fact somewhat irrelevant to the second part. Studies in predictive processing have paved the way for this top-down framework and there is no obvious reason to refute a similar framework in theories of consciousness. As Metzinger (2003: 52) also says,

> [A] fruitful way of looking at the human brain, therefore, is as a system which, even in ordinary waking states, constantly hallucinates at the world, as a system that constantly lets its internal autonomous simulational dynamics collide with the ongoing flow of sensory input, vigorously dreaming at the world and thereby generating the content of phenomenal experience.

Undoubtedly further work on how HOT theory of phenomenal consciousness and PPT can be brought together is needed and for reasons discussed this seems to be a promising way to enhance our understanding of the mind and of consciousness.

References

Berger, J. 2014. "Consciousness is not a property of states: A reply to Wilberg." *Philosophical Psychology* 27 (6): 829–842.

Block, N. 2011a. "The higher-order approach to consciousness is defunct." *Analysis* 71 (3): 419–431.

Block, N. 2011b. "Response to Rosenthal and Weisberg." *Analysis* 71 (3): 443–448.

Byrne, A. 1997. "Some like it HOT: Consciousness and higher order thoughts." *Philosophical Studies* 86 (2): 103–129.

Clark, A. 2016. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford: Oxford University Press.

Farrell, J. 2017. "Higher-order theories of consciousness and what-it-is-like-ness." *Philosophical Studies* 175 (11): 2743–2761.

Gennaro, R. J. (ed.). 2004. *Higher-Order Theories of Consciousness: An anthology*. Amsterdam: John Benjamins Publishing.

Gennaro, R.J. 2012. *Consciousness Paradox.* Cambridge: MIT Press.

Gennaro, R. J. 2017. *Consciousness.* New York: Routledge.

[9] Unless the HOT itself is conscious which is rarely the case, if possible.

Hohwy, J. 2013. *The Predictive Mind.* Oxford: Oxford University Press.

Jackson, F. 1986. "What Mary didn't know." *Journal of Philosophy* 83 (5): 291–295.

Kriegel, U. 2003. "Consciousness as intransitive self-consciousness: Two views and an argument." *Canadian Journal of Philosophy* 33: 103–132.

Kriegel, U. 2004. "Consciousness and self-consciousness." *The Monist* 87 (2): 182–205.

Levine, J. 2001. *Purple Haze: The Puzzle of Conscious Experience.* Cambridge: MIT Press.

Neander, K. 1998. "The division of phenomenal labor: A problem for representational theories of consciousness." *Philosophical Perspectives* 12: 411–434.

Mandik, P. 2009. "Beware of the unicorn: Consciousness as being represented and other things that don't exist." *Journal of Consciousness Studies* 16 (1): 5–36.

Metzinger, T. 2003. *Being No One: The Self-Model Theory of Subjectivity.* Cambridge: MIT Press.

Metzinger, T. and Wiese W. (eds.) 2017. *Philosophy and Predictive Processing.* MIND Group.

Rosenthal, D. 1991. "The Independence of Consciousness and Sensory Quality." *Philosophical Issues* 1: 15–36.

Rosenthal, D. 1993a. "Higher order thoughts and the appendage theory of consciousness." *Philosophical Psychology* 6 (2): 155–166.

Rosenthal, D. 1993b. "State consciousness and transitive consciousness." *Consciousness and cognition* 2 (4): 355–363.

Rosenthal, D. 2000. "Metacognition and higher order thoughts." *Consciousness and cognition* 9: 231–242.

Rosenthal, D. 2002. "Explaining consciousness." In D. Chalmers (ed.). *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: Oxford University Press, 406–421.

Rosenthal, D. 2004. "Varieties of higher order theory." In R. Gennaro (ed.). *Higher-Order Theories of Consciousness: An anthology.* Amsterdam: John Benjamins Publishing, 17–44.

Rosenthal, D. 2005. *Consciousness and Mind.* Oxford: Oxford University Press.

Rosenthal, D. 2011. "Exaggerated reports: A reply to Block." *Analysis* 71 (3): 431–437.

Weisberg, J. 2011a. "Misrepresenting consciousness." *Philosophical Studies* 154: 409–433.

Weisberg, J. 2011b. "Abusing the notion of what-it-is-like-ness: A response to Block." *Analysis* 71 (3): 443–448.

Wilberg, J. 2010. "Consciousness and false HOTs." *Philosophical Psychology* 23 (5): 617–638.

Croatian Journal of Philosophy
Vol. XXII, No. 64, 2022
https://doi.org/10.52685/cjp.22.64.6
Received: November 20, 2021
Accepted: April 29, 2022

Non-Stupidity Condition and Pragmatics in Artificial Intelligence

BOJAN BORSTNER and NIKO ŠETAR
University of Maribor, Maribor, Slovenia

Symbol Grounding Problem (SGP) (Harnad 1990) is commonly considered one of the central challenges in the philosophy of artificial intelligence as its resolution is deemed necessary for bridging the gap between simple data processing and understanding of meaning and language. SGP has been addressed on numerous occasions with varying results, all resolution attempts having been severely, but for the most part justifiably, restricted by the Zero Semantic Commitment Condition (Taddeo and Floridi 2005). A further condition that demands explanatory power in terms of machine-to-human communication is the Non-Stupidity Condition (Bringsjord 2013) that demands an SG approach to be able to account for plausibility of higher-level language use and understanding, such as pragmatics. In this article, we undertake the endeavour of attempting to explain how merging certain early requirements for SG, such as embodiment, environmental interaction (Ziemke 1998), and compliance with the Z-Condition with symbol emergence (Sun 2000; Tangiuchi et al. 2016, etc.) rather than direct attempts at symbol grounding can help emulate human language acquisition (Vogt 2004; Cowley 2007). Along with the presumption that mind and language are both symbolic (Fodor 1980) and computational (Chomsky 2017), we argue that some rather abstract aspects of language can be logically formalised and finally, that this melange of approaches can yield the explanatory power necessary to satisfy the Non-Stupidity Condition without breaking any previous conditions.

Keywords: Artificial intelligence; symbol grounding; pragmatics; language; computationalism.

1. *Introduction*

Artificial intelligence is as hot a topic as any during the last few decades, with debates on it ranging from AI ethics to its development to whether it is achievable at all. Currently, a lot of progress is being made in the development and production of neural networks and machine learning systems, yet it would seem that those systems are still not much more than just increasingly sophisticated software on increasingly sophisticated hardware. The key difference between them and artificial intelligence is, well, intelligence. Here we reach a whole different debate: what exactly does it mean to be intelligent? There is an abundance of answers, or at least attempts at answering, but let us make it simple and agree that intelligence is inextricably linked with understanding – therefore, in order to be intelligent, a machine has actually to understand the data it is processing, and not just merely process it. And that is only the beginning in the long process aimed at achieving human-like intelligence or even superintelligence.

In this article, we will overview one of philosophers' favourite approaches to making AIs understand their data – solving the Symbol Grounding Problem, which we shall introduce in the next section. We will study several proposed solutions, cherry-picking certain elements to comprise a strategy with a decent chance of success. Afterwards we will address the issue of whether any approach to grounding has the explanatory power as to how human-level artificial intelligence could be achieved and explain how this may be within our reach if we explain how complex features of language such as speech acts, metaphors, and humour may be grounded in simpler features (non-connoted sentences, words) that are in turn grounded directly.

2. *Symbol grounding problem*

The Symbol Grounding Problem was first formulated by Stevan Harnad (1990) and is derived from John Searle's (1980) Chinese Room thought experiment. Searle describes a room containing a vast number of monolingual resources in Mandarin Chinese, from dictionaries to encyclopaedias and novels. There is also an English-speaking man in this room who has no knowledge whatsoever of the Chinese language or writing. Next, we insert a paper page with a number of questions in Chinese that our man in the room must answer. Searle claims that with enough time (or processing power) he can find corresponding patterns of symbols in the available resources and copy the symbols that follow the question mark until the end of the sentence or paragraph. Then he outputs the paper with what are likely perfectly correct answers. However, through this process, the man in the room never understood a single Chinese symbol he was looking up or copying and had no idea what the input questions or his own output answers were. This is analogous to how computers process data: they operate based on an algorithmic script. When they receive an input X, they 'look for' a part of their code that says something like "if X then Y," and output Y accordingly. Thus, when a command is typed into a computer and the computer performs this command, it does so without understanding what it just did, what the input meant, or what the output meant.

Harnad (1990) says that symbol grounding problem comes in two forms, the first of which is not unlike learning Chinese as a second language, using only a monolingual Chinese dictionary, which is a rather difficult task. The second form is like trying to learn Chinese as one's first language using only such a dictionary – an impossible task. Since symbol grounding that we are talking about when speaking of AI is essentially a form intended to ground a first language, such learning is impossible. What we need are external (real-world) referents to which we can relate the symbols we are manipulating.

2.1. *Approaches and conditions*

Harnad (1990) himself proposes a representationalist approach towards symbol grounding. The approach is based on the notion that the distal objects are projected onto the perceiver's sensory surfaces when they are perceived via any available means, be it sight, hearing, touch, or any other sensory tool, and is drawn from the work of Shepherd and Cooper (1982). Harnad dubs these projections as representations and defines several kinds of representations that manifest throughout the process of transcription of distal objects into symbols within one's mind as a symbolic system (for further details on mind as a symbolic system see Fodor 1980). When we are exposed to a particular referent in the outside world, an iconic representation of it is created; a group of referents with similar properties, in turn, yields categorical representation. Two cognitive mechanisms manipulate those representations: discrimination allows us to distinguish between different categories, as well as different tokens within a category; identification lets us recognise something in the outside world as a token belonging to a category. When related to a particular symbol (spoken/written word or such), the symbolic representation of a token or a category is formed. Regier (1992) attempts to recreate a similar bottom-up procedure by taking artificial agents equipped with cameras and presenting them with a number of photographs, which served as a base for him to teach them some basic two-dimensional spatial relations – the experiment was not entirely unsuccessful, but it seems apparent that it achieved only basic machine learning rather than grounding.

The approaches above are both cognitivistic; that is to say they belong in the group of approaches to various mind-related problems that distance themselves from agents' behaviours and rather focus on underlying processes within the mind that elicit said behaviours. However, neither of them yielded desired results, which some saw as bad news for cognitivism in symbol grounding in general. Ziemke (1998) argues that this is because they are simply tagging things out there with prescribed symbols, and do not interact with them enough to be able to achieve grounding. Ziemke therefore proposes what he calls enactive grounding. This approach is based on true interaction between the artificial agent and its environment, which calls for agent embodiment, i.e., the agent must be given a physical form that allows it to interact as much as possible, therefore including visual, audio, and any other possible receptors. With such a system, it is also possible to arrive at behaviour emergence (a behaviour emergent from agent's interactions, independent from its source code or such), and, by extension, grounding emergence. Another example of an enactive approach is Sun's (2000) approach, which mainly relies on phenomenology, claiming that an agent has to be embodied to be in the world and to be able to make itself available for recognition in the world. Both of these enactive approaches are facing the externalist trap, which is the reduction of agent's behaviour to mere reactions to external factors in its environment, placing the environment first. If all behaviour of the agent is nothing but a reaction to outside stimuli, then the agent cannot be considered autonomous (Ziemke 1998). This is part of a more fundamental question of how exactly an artificial agent and the environment would interact, beyond the AI simply recording the environment and again, merely tagging things with symbols.

Enactivism has generally proven to be a rather popular approach within cognitive science and can be primarily described as a position that seeks to explain cognition and mental processes as a complex set of interactions between a living agent, its immediate environment, and the world in general (Varela, Thompson and Rosch 1991). According to these authors, enaction itself is the process in which a perceiving agent acts (either consciously or automatically) to the requirements of its environment and given situation. This basic form of enactivism is known as autopoietic enactivism, where autopoiesis refers to the process of self-maintenance and autonomy It is supposed to both present an alternative to dualism in the sense that the distinction between mental and biological processes is almost eliminated, and the former seem to supervene on the latter, as well as distance itself from representationalism (Maturana and Varela 1992).

The notion of this distancing is better shown within the theory of sensorimotor enactivism, which claims that perception is an active, rather than passive process, where perceiving agents actively explore and intentionally seek to interact with the world. In those interactions, they appeal to sensorimotor expectations about how objects in the world will change depending on the agent's angle of perception, physical interactions with said objects, etc. (Noe 2004). These expectations are what then defines cognition, and are considered to be non-representational, although it could be argued that they still demand some degree of mental modelling of the expected states of the world.

Finally, theories of radical enactivism seek to eliminate representation altogether. Hutto and Myin (2013) for example go to great lengths to deconstruct various preceding views of cognition, including those found in autopoietic and sensorimotor enactivism, in order to explain them purely in terms of enaction and without any need for representation. Surprisingly, they arrive at the conclusion that representations can be avoided only on the level of basic cognitive and perceptual processes, i.e., when dealing with concrete objects and concepts, and that complex processes such as language nevertheless need to rely on representations to process abstractions and symbols in language.

These enactive approaches are therefore all still based in representationalism. Although they seek to distance themselves from representationalism, autopoietic approaches never claim they have done so entirely, sensorimotor approaches seem to revert to them at least partially when one considers how exactly "expectations" are manifested in the agent, and radical approach admits it is only possible on rather basic levels. Theories of enacted cognition have great potential in pursuit of grounding in artificial agents as they complement embodied cognition remarkably well, as well as present an adequate basis for (symbol) emergence, which we will mention later. Now, however, we shall return to our analysis of other various approaches.

Next to be explored is the functional model developed by Mayo (2003), where what Harnad considers categorical representations are interpreted in a functionalist sense. Every category is considered a set that contains functionally relevant elements. A single symbol may evidently therefore exist in several functional categories. Mayo claims that it is this very overlap in functions of one discrete symbol that characterises it as distinct from those who share some but not all its functions. These various representationalist approaches are important because newer approaches to the symbol grounding problem tend to return to representationalism at least in the early stages of the grounding procedure. Still, we shall briefly mention semi-representationalist and non-representationalist approaches as well.

One of the semi-representationalist is, for example, the physical symbol grounding problem, where a symbol is considered a physical form of what is represented. A semiotic symbol system consisting of form, meaning, and referent, is introduced; in that, the form is the physical tag of a symbol, the meaning the semantic content of the physical tag, and the referent is the "thing" in the outside world to which the tag applies. Artificial agents then attempt grounding through an imitation game consisting of speaker agents and hearer agents. The speaker agents vocally express the symbolic tag of the referent, while the hearer agents must figure out what it applies to. The idea is that the symbol (symbolic tag) is grounded in the hearer agent when it can accurately recognise the referent upon hearing the tag (without intermittent mistakes) (Vogt 2002). Finally, non-representationalist models entirely disregard representations' role in the symbol grounding problem and instead fully rely on the interaction between the artificial agent and its environment.

A breakthrough is made by Taddeo and Floridi (2005), who review all significant research on the topic since Harnad, finding that none of the approaches above, as well as numerous others we left out in this analysis, satisfies what they call the Zero semantic commitment condition or Z-condition for short. The latter is formalised as follows:

> 1) No form of *innatism* is allowed; no semantic resources (some *virtus semantica*) should be presupposed as already pre-installed in the AA; and
> 2) no form of *externalism* is allowed either; no semantic resources should be uploaded form the "outside" by some *deus ex machina* already semantically-proficient.
> Of course, points (a)-(b) do not exclude the possibility that
> 3) the AA should have its own capacities and resources (e.g., computational, syntactical, procedural, perceptual, educational etc., exploited through algorithms, sensors, actuators etc.) to be able to ground its symbols. (Taddeo and Floridi 2005: 423)

Most forms of approaches we described above rely on innatisms, which are indeed problematic for symbol grounding, but some merely rely on certain externalisms, that we will later argue can be sometimes justified in analogy to human grounding.

The same authors later (2007) establish their approach to symbol grounding that they claim satisfies the Z-condition and brings one as close as possible to solving the problem in question. The first principle they introduce is the Action-Based Semantics, which assumes that meanings are in their first stage internal states of the agent, whereafter they trigger actions, which proves them to cause semantic emergence in the agent (without innatism). The second principle is the division of the agent into two machines that both communicate with the environment and each other, thereby allowing the agent to reflect on its actions. This latter principle allows access to communication capacities, categorisation/abstraction capacities, and representational capacities within the agent, as well as access to feedback. The former principle provides a sensomotorical interactive approach, as well as an evolutionary approach and the satisfaction of Z-condition.

As successful as this approach may seem, Bringsjord (2014) emphasises that Taddeo and Floridi's approach lacks the explanatory power as to how an artificial agent, functioning based on their design, could reach the level of grounding where it could communicate on the same level as a competent human speaker. Bringsjord invokes an example of a letter written by a girl to her boyfriend, which a human reader (such as me or you) can plainly understand to be sarcasm; a good approach to grounding must be able to explain how an artificial agent can reach the level of understanding sarcasm, humour, pragmatics, metaphors, etc. Bringsjord himself notices that Z-condition might be blocking that possibility entirely on higher levels of grounding, while evolutionary approach to grounding also seems to be quite faulty.

The issue with the evolutionary approach is that there is no concrete evidence that human linguistic competence developed strictly through evolution since some early linguistic features were quite redundant as per humans' needs at the time (Bringsjord 2014); it is also hard to grasp how simulating the entirety of human language evolution in an individual artificial agent would make any sense. As Harnad (1990) implies in the Chinese Merry-go-round description, an artificial intelligence attempting symbol grounding is not unlike a baby learning its first language, and by no means does a baby lying in her crib have to invent words for things she sees around here. She will not replicate linguistic evolution and emerge at 18 months old with a private language, rather, she will learn the language(s) of her parents by interacting with them and their environment, and it is likely this principle of human language development we should follow when pursuing symbol grounding.

2.2. *Human grounding simulation*

The first thing that seems to be quite on point about this notion is that it is evident that children learn their first language – for which they have to acquire symbol grounding – through interaction with their environment (Vogt 2007). The children learn their first language by attributing meanings to symbols depending on the symbols' context in terms of both other symbols as well as perception data available (e.g., if someone is pointing at a particular thing when uttering a symbol). The agent must decide on a symbol's meaning depending on all of its contextual features. Vogt serves an example where a linguist hears a native speaker of an unknown language utter "Gavagai" when a rabbit appears on the scene. Purely logically, the auditory symbol "gavagai" could mean numerous things, but for humans it is intuitively very easy to determine its most likely meaning is "rabbit." We may remark here that the original use of the Gavagai example appears in Quine (1960), where the linguist in question undergoes a tedious procedure of verifying her assumption that "gavagai" is more likely to mean "rabbit" than "white" or merely "animal" by studying the natives' affirmative and negative responses to her using "gavagai" in those varying contexts. Our point here, however, relates to none of these. Rather, what we wish to take away from this example is how easy it is for humans to intuitively grasp the most likely meaning of a new word, immediately favouring the more likely "rabbit" over less likely but plausible "white" or "animal."

An artificial agent, however, may have trouble recognising instances on its own, therefore it would likely require some prerequisite competencies that would allow it to be able to make such a connection as the one between "gavagai" and a rabbit. It should, for instance, somehow know what it means when somebody points at something, as humans seem to intuitively even at a very young age. We are, again, not claiming that humans can determine with utmost certainty the meaning of any new word; we are simply observing that we seem to have a predisposition to pick out the most likely of various possible meanings with a decent degree of success.

Cowley (2007) offers a solution to this dilemma when he describes that a (human) baby primarily relies on the role of its parents when learning to communicate. Namely, it relies on the notion that its parents will demonstrate, by communicating to it and each other, an appropriate pattern of actions, vocalisations, and relations between action and vocalisation. What happens in this procedure is that children learn to speak by being explained or shown symbols their parents have already grounded. Children finally become competent speakers by coordinating with the others consistently in a certain cultural or social environment. In reference back to Quine and Vogt's Gavagai example, a child gets to know the meaning of "rabbit" from being shown a rabbit (or an image thereof) by her parents, accompanied by them uttering the word "rabbit," presuming they know what a rabbit is and that the symbol "rabbit" refers to that particular fluffy creature. In a later circumstance, the same child, now adult, will assume (likely correctly) that "gavagai" means "rabbit" rather than "white", because that is how her parents demonstrated new symbols. Of course, in this later context, Quine's verification procedure applies, as it does for artificial agents, which we will show later, noting also that for artificial agents all possible meanings of a symbol carry the same probability value, which is not true for human agents. For Cowley, there is also no pure symbol as far as humans are concerned – rather, symbols are a posteriori and derived from the use of language, grounded in behaviour and action.

Another type of simulation that we may require to achieve grounded cognition is a more direct simulation of cognition itself (Barsalou 1999, 2008; Pezzulo et al. 2013). Barsalou (1999) proposes an approach named Perceptual Symbol Systems theory, which acknowledges that modal symbolic operations are of great importance for interpreting experience and suggests that natural implementation of such operations can be achieved by the means of mental simulations. According to the theory in question, there is "a single, multimodal representation system in the brain that supports diverse forms of simulation across different cognitive processes" (Barsalou 2008). Such cognitive processes include several types of perception, various levels of memory, as well as conceptual knowledge. This allows for (multimodal) states to be captured in memory and retrieved to be simulated when required. These processes occur within human cognition, as well as, according to Barsalou, in non-human agents (in this case, animals). Reasonable assumption is that such systems of mental simulation should also be computationally emulated within artificial agents to achieve grounded cognition and in turn symbol grounding (Pezzulo et al. 2013).

Barsalou (2008) emphasises on the link between language and simulation, pointing out several examples: situation models, perceptual simulation, motor simulation, affective simulation, and gestures. Situation models are spatial representations, or better yet, spatial situation simulations that occur when scenes from written texts are described to people verbally, showing a tight relation between visual and verbal comprehension of spatial situations. Perceptual simulations refer to the representations an agent constructs when a concrete object is described to them; when a description of an object is vague, the representations contain implicit perceptual information about the object, which is more than likely drawn from the agent's memory. Next, motor simulations occur when verbs for actions of various body parts are described to the agent, which triggers a reaction in their motor system; according to Barsalou (2008) neurological research had shown this happens on the level of the central nervous system even when the corresponding action does not manifest physically. Fourth, affective simulations are those that occur when an agent is exposed to a word, or a text, that carries some form of emotional value for the agent. Finally, gestures are an expression of embodiment in language that connect bodily movements with the meanings of words they accompany. Barsalou (2008) provides numerous examples from empirical studies that support all of the above types of simulation-language relations. However, such examples are hardly in the scope of this paper, but we encourage the reader to refer to the original text by Barsalou.

Grounded cognition through mental simulations as summarised above can greatly contribute to achieving symbol grounding; a great additional illustration of this can be found in Pezzulo et al. (2013) where the authors explain the "cascade of effects on cognition" from grounding through embodiment to situatedness. It also concurs with the requirement for human-grounding simulation we have discussed at the beginning of this section (in Vogt 2007 and Cowley 2007), as well as with requirements for multimodality and embodiment (e.g. Ziemke 1998; Cangelosi and Riga 2006).

We would like to pause to address the issue we mentioned with Taddeo and Floridi's Z-condition. Particularly that the second point of their condition, which prohibits any and all kinds of externalism is too stringent. If we look back to Cowley, we see that children seem to learn at least in part by being explained symbols by agents who are already semantically proficient, that is to say they have already grounded those symbols. A simple example of this would possibly be a child's mother pointing at herself and saying "momma" when interacting with her toddler. Eventually, every healthy child will successfully learn that "momma" is that female figure that feeds her, consoles her, plays with her, etc., and learn to point at her and say "momma" as well. At later stages, the child may be attending school, where she is very plainly explained the meaning of the word "addition" in mathematics or "gravity" in physics. If such externalist explanations do not violate human grounding process, why should they be considered as violations of artificial agents' grounding processes? Indeed, without such externalism we seem to be forever stuck on a version Harnad's impossible version of the Chinese Merry-go-round where we expect an agent to learn a first language from a dictionary. To prevent such conundrums, certain externalisms have to be allowed in the grounding process. Of course, the process should not be fully reliant on them, as children learn plenty by simply observing what others vocalise in different contexts and learn to replicate that quite successfully on their own.

This can greatly contribute to what is already known in robotics as the epigenetic model and can feature in Emergent symbol grounding approaches (Tangiuchi et al., 2016). The latter proposes that in humans, symbols emerge throughout the language learning process, wherein they automatically connect to referents and each other, thereby grounding themselves in perceptions, internal representations of those perceptions, and actions. Tangiuchi et al. introduce their own requirements for this model to be successful. One of those is, for instance, multimodal categorisation, which requires agents to ground every category (of things) in multiple modalities, i.e., visual perception, audio perception, haptic perception, and any others available. Thus grounded (categorical) symbol includes all perceivable features of the thing or all common perceivable features of the category of things in which it is grounded.

An interesting example of an early epigenetic model is Cangelosi and Riga's (2006) experimental embodied agent. They suppose two grounding mechanisms: the first grounds basic vocabulary directly in environmental interaction; the second one is transferred grounding that allows the agent to join two basic grounded elements and ground in them a more complex symbol. We will not go into many details of the experiment. The robots had a number sensomotoric actions in their programming but lacked any symbol to connect them with – upon receiving a symbolic order, such as "Close left arm," they randomly performed one of those actions and received positive feedback if right. The first phase consisted of repeating this procedure on several basic phrases. The second phase contained phrases such as "Grab" and the agents had to "figure out" that "Grab" consists of "Close left arm and Close right arm." In the third phase, they had to ground phrases that were conjunctions of the second phase phrases. The experiment was rather successful with a high rate of accuracy on all three stages; however, even the basic stage required a large number of repetitions, with the second and third requiring respectively more. This can be partially ascribed to the processing power of computers fifteen years ago, or we can say that perhaps symbol grounding is a procedure that is just as long and complex as first language learning is in children.

What have we ended up with at this point? It seems like that in order to achieve grounding, we require:

1. An embodied agent with multimodal capacity

2. An epigenetic approach to symbol grounding (simulating human first language acquisition and human cognition in terms of mental simulations)
3. A multi-phased approach to symbol grounding (allowing complex symbols to be grounded in baser symbols, or to be simply explained)
4. For the purposes of 2 and 3: dropping the second requirement of the Z-condition
5. To be prepared the procedure may take a very long time (as a consequence of 2)
6. Explanatory Power for the Non-stupidity Condition

It is this last point the second half of our article will focus on.

3. *Explanatory power for pragmatics*

If we are to move on to satisfying the Non-stupidity condition, the first thing we ought to do is explain how grounding abstract symbols can be achieved as some nth phase of our multi-phase grounding model, wherein the early phases involve grounding very concrete, physical symbols with increasing complexity. What we consider an abstract symbol is a symbol without a physical or directly perceivable (by means of multimodal sensory apparatus) referent in the outside world (Cangelosi and Riga 2006; Šetar 2020b; Tangiuchi et al. 2019)

Initially, some basic symbol grounding is quite correctly described already by Harnad, albeit in a representationalist way. Harnad claims that once we have grounded both the symbol "horse" and the symbol "stripes" – in this case we are grounding them nicely and slowly through epigenetic, multimodal interaction – we can ground the term "zebra" without actually having any experience with the primary referent for "zebra." It is enough that an agent experiences pictures or films of a zebra but can also form an idea of a zebra as a black-and-white striped horse similarly as "horn" and "horse" can lead to the idea of a unicorn. However, these sorts of conjunctions only seem to function as far as concrete symbols with physical referents are concerned.

To understand how grounding might proceed for abstract concepts and pragmatic elements, we can look at four requirements proposed by Tangiuchi et al. (2019):

> – Creating holistic language processing systems that involve physical, psychological, social, conceptual, and experiential constraints.
>
> – Inventing machine learning methods to represent the recursive property of background beliefs for holistic language processing.
>
> – Developing computational models for collaborative tasks in the physical world, leading to the emergence of dialogue.
>
> – Inventing methods to enable a robot to make use of contexts, e.g., situation and culture, and to grow the ability to use language to exchange meaning by referring to social factors: field, tenor, and mode. (Tangiuchi et al. 2019: 20)

While developing computational models is a matter best addressed by those with more technological prowess than the authors of this article, and inventing machine learning methods, even just theoretically, is a detailed and tedious task that falls out of the scope of this article, the first, and especially the latter requirement may shed some light on the issues at hand. Tangiuchi et al. (2019) look for a solution in Halliday's functional linguistics, where the semantics of a word depends on its contextual use, depending on culture and particular situation. And while situational and cultural contexts may be taught to artificial agents with some additional effort, we shall seek a solution elsewhere – namely, Chomsky's theory of universal grammar (1957). The idea we are focusing on here is that every sentence has a kernel unit, while the sentence is a transformation of that kernel. The transformation itself is not a matter of semantics but rather a tool for the disambiguation of meaning based on socially defined functional semantics. This offers us an option to pre-equip our artificial agent with a non-semantic grammatical apparatus that enables syntactic formation and transformations; the latter are defined by the interaction of our agent with its environment, which teaches it, by providing examples to be analysed, which transformation is correct in what context. The sentence kernels are those symbols that need to be grounded in the traditional sense. Additional insight is offered in more recent Chomsky (2017), where the author determines that given the speed at which language is acquired by children and the low amount of presentation required for them to learn and ground a new linguistic symbol, language itself or at least the basis thereof must be deeply internalistic and supervene on simple computational processes, with all externalisms coming in later, allowing for communicative faculties of language. While some other aspects of the article in question pose some new issues for language grounding in artificial intelligence, mainly in the environmental interactivity department, there is an important new point to be made. If language is, when sufficiently reduced to its evolutionary core, indeed a simple computational process, then this computational process may be quite easily replicated in artificial neural networks once it is determined how it works on a formal computational level in humans. The notion that the (generative) acquisition of one's first language is deeply internalistic and requires very few presentations, also entails that the internalist trap (the opposite of the externalist trap defined earlier) is not in fact a trap, but a necessary first step in language development. Referring to Vogt and Quine's example, we learn the word "rabbit" via a computational, internalist process that pertains to acquiring one's first language, and later affirm it and attempt to disambiguate "gavagai" in virtue of second-order externalist processes that pertain to effective communicational use of our first language as well as acquiring further languages. Much further ado is necessary here, which would only confuse the rest of this article, but may serve as a basis for an entirely separate one in the future.

Going back to allowed preconditions for symbol grounding – field, tenor, and mode: all of the elements are in and of themselves non-semantic and could therefore be used as a tool in an epigenetic model for symbol grounding. However, the field requires an understanding of topics, and cultural and social context, which can only be learned through interaction and communication; therefore, it cannot be precluded in an agent. We have similar issues with mode, which characterises discourse structure, way of expression, etc; again, slangs, registers and such must be learned as part of satisfying the Non-stupidity Condition. Lastly, however, some parts of tenor may be precluded in a learning agent. While it will develop social relations with other agents on its own, it is in no contradiction with epigenetic modelling to pre-equip an artificial agent with devices that allow it to perceive certain tones of voice, pitches, etc. as negative or positive, seen as a baby has no issue distinguishing between, for example, a parent being upset and a parent being caring.

Another concept that may be required to proceed from concrete symbol to abstract symbol grounding is the concept of semantic affordance (Glenberg and Robertson 2000). A chair, which can basically be defined as a piece of furniture with four legs affords humans with a function of sitting but does not afford the same function to an elephant, while it affords this function to a cat only incidentally but not intentionally. There are also contingent affordances, such as the affording the function of being stood on to reach a higher location.

It is multimodal sensory experience that first helps ground the notion of "chair" and it also helps extend this notion to a variety of chairs – those with three legs, those without a back, etc. It is at a later stage that "leg [of a chair]" is grounded as part of a chair and distinctly from "leg [of a human]." However, "chair" is a very simple, concrete symbol, and so is "leg [of a chair]," even though it is located a phase higher in grounding hierarchy than "chair."

Finally, let us look at how one could ground "[a] painting." In the earliest multimodal grounding phase, we would need an experience of seeing a number of depictions of things, which are not photographs and not printed in any other form; haptic perception (i.e., touch) could be of help here in recognising the texture of a painting. Next, we would need to have already grounded concepts of "form [in general]" and "content [in general]," which an agent would then have to specialise to "form [in painting]" and "content [of a painting]" – this can be done by explaining the agent how these concepts work in painting just as an art teacher would explain it to students. Several stages later, a complex grounded scheme like "(if 'form' is 'dynamic'… and 'content' is 'exaggerated,' 'twisted'…)" can mean "expressionism." These notions are extremely difficult to describe in humans, not to mention in artificial agents. The point is, however, that in humans such multi-layered approach to grounding evermore complex and abstract symbols seems to work – therefore, why should it not in a sophisticated epigenetic artificial agent?

3.1. *In speech acts*

While speech acts were first formulated by Austin (1962), we will not use his threefold classification (locution, illocution and perlocution) in our attempt to describe possible grounding mechanism for speech acts because we argue (Šetar 2020a and b) that locution, illocution and perlocution are in fact features of speech acts that every speech act possesses.

Instead, we will use a more contemporary classification of speech acts into the assertive, commissive, constative, directive, and imperative speech acts (Jary 2010; Kissine 2013; Jary and Kissine 2014). Assertive speech acts are statements that are truth-bearing and convey truth-value information without explicit intention of altering the hearer's belief; commissive speech acts are ones that speaker uses to commit themselves to fulfil their content, such as promises and threats; constative speech acts are ones intended to alter the hearer's belief regardless of their de facto truth value; directive speech acts intend to convince the hearer to fulfil their content by providing sufficient reason to do so; lastly, imperative speech acts instruct the hearer to fulfil their content without providing a reason but rather do so by other means, most commonly by being uttered from a position of authority. The five classes of speech acts can be formalised as follows:

> Assertive: "A is and assertive speech act containing proposition p if, and only if, the speaker believes p to be true and there is justification for p to be true." (Šetar 2020a: 35, drawing on Jary 2010)
>
> Commissive: "All promises are acts of placing oneself under an obligation to bring about the propositional content p." (Kissine 2013: 149)
>
> Constative: "An utterance is a constative speech act with the content p if, and only if, with respect to this background, it constitutes a reason to believe that p." (Kissine 2013: 62)
>
> Directive: "An utterance is a directive speech act with the content p if, and only if, with respect to a given background, it constitutes a reason to bring about the propositional content of p." (Šetar 2020a: 44, drawing on Kissine 2013)
>
> Imperative: "I is an imperative speech act containing proposition p if, and only if, it compels the hearer to bring about the propositional content of p." (Šetar 2020a: 46, drawing on Jary and Kissine 2014)

But why do we require such formalisations in the first place? That is due to the fact that humans recognise the function and intention of speech acts entirely intuitively, that when hearing a certain phrase, we do not have to break it down and consciously consider what speech act it is, we simply know. This could be an inherent faculty of ours being conscious, and since it would be terribly reductive for one to assume that symbol grounding or any other form of artificial intelligence entails consciousness (see Pierce 2017), we must find a mechanism to teach speech acts to an agent that is not necessarily conscious and does not necessarily possess intuitions or other such capabilities. Given the logical nature of programming and computer operations, logical formalisations of speech acts are a reasonable way out. However, we need a concrete symbolic referent through which a speech act can be determined to belong to a certain class. In Šetar 2020a we found that a viable candidate for this in English may be modal verbs, which can also be nicely logically formalised:

> Can: p is compatible with the set of all propositions which have a bearing on p.
> May: there is at least some set of propositions such that p is compatible with it.
> Must: p is entailed by the set of all propositions which have a bearing on p.
> Should: there is at least some set of propositions such that p is entailed by it. (Where p is the proposition expressed by the rest of the utterance). (Papafragou 1998: 50)

Modals "have to" and "ought" to be also considered here; for the purposes of this article "have to" is seen as an equivalent of "must", and "ought" is formalised between "must" and "should," as it is generally perceived as deontically weaker than "must", yet stronger than "should". We explain this in a bit more detail in Šetar (2020a), where we draw on Groefsema's (1995)'s formalisations of modals "must" and "should," also summarised in Papafragou (1998). If in "must," the contained proposition p is entailed by all prepositions that have a bearing on it, and in "should" it only needs to be entailed by some arbitrarily small set of such propositions, we can say that in "ought," p is entailed by most of the propositions which have a bearing on p.

What this brings us is the notion that assertive speech acts can be those that are either non-modalized or involve entailing modals "have to," "must" and "can" in an epistemic sense, which is to say they convey a certain knowledge or belief. The need for strong entailing modals arises from the fact that assertive speech acts necessarily convey knowledge and not mere belief.

Unlike assertive speech acts, constative speech acts are ones intended to convince the hearer of speaker's belief (not necessarily knowledge), they can feature any modal used in an epistemic sense. For example, "there should be a connection between those events" is a constative speech act, and so is "they must be brothers." However, "increasing summer temperatures must be related to global climate change" is an assertive speech act.

For commissive speech acts we can say they are those using "must" and "have to", as well as sometimes "ought to" in first person, in a deontic way – the latter meaning that they express a duty to do something: specifically, to bring about the proposition contained in the utterance. "Will" can also be considered a modal verb that shows intention to do something and can therefore also be an indicator of a commissive speech act.

Directive speech acts are also based on deontic use of modals and, like assertive speech acts, require entailing modals, albeit not only the stronger ones. Thus, "you must finish your homework" and "you should not be late again" are both directive speech acts. They can, however, also be imperative, depending on what kind of deontic justification lies behind their use. If the former is spoken by a teacher and the latter by a boss, they are justified by authority and therefore certainly imperative – yet if they are uttered by the hearer's friend they are directive, as they are otherwise justified, for example as "you must finish your homework [if you wish to pass the course]" and "you should not be later again [if you wish to avoid disciplinary action]".

It is reasonable to also mention performative speech acts, which are difficult to formalise in the way presented above, as they are speech acts that alter something in social (conventional) reality, if uttered from a position of proper authority. Notable examples are "I now pronounce you man and wife" as uttered by a priest, or a parent naming their new-born child.

Even though a modal verb can be an excellent cue for the artificial agent to start identifying a speech act as belonging to a certain class and having a certain function, it does not fully define a speech act. What is still necessary is for the artificial agent to have certain conception of epistemic and deontic use, as well as of authority. This is where we refer back to the emulation of grounding development in humans and pre-given capabilities related to recognising tone, mode, and field of discourse. An artificial agent with a long enough learning process will have grounded the concept of authority relatively early in that process and will be able to distinguish different uses of the same modal verb depending on the pattern of their use by others. That is to say that it should be able to conceive of "you must clean this room" as imperative or directive based on the deontic "must", while also being able to understand that "you must try these cookies" is in no way an imperative or even a directive, based its interaction with environment, i.e. based on how "must" is usually used, in what contexts it is used, and how human hearers react to it depending on its uses in different contexts.

3.2. *In metaphors*

An important aspect of satisfying the condition of non-stupidity is accounting for how metaphorical speech may be grounded since that very type of speech is commonplace in everyday communication in idioms, proverbs, literature, etc. In doing so we will first refer to the notion that metaphorical utterances can be understood in two ways: through their original domain or through the target domain (Tangiuchi et al. 2019). The original domain involves concrete symbols and concepts whose referents are usually empirically accessible, i.e., the literal meaning of the phrase, while the target domain is the translation of those symbols and concepts into their abstract meaning, which is semantically related to the literal meanings in the original domain.

Let us examine the idiom "she wouldn't harm a fly." If this idiom was to be understood in context of its original domain it would be interpreted as if the person in question has an actual, literal aversion towards harming a particular type of insect. That sort of interpretation is certainly stupid in Bringsjord's sense. In context of its target domain, however, it means that the person to whom the metaphor refers is very peaceful and gentle. Where that derives from is the conception that striking a buzzing fly is generally considered an extremely mild, or even the mildest conceivable form of violence. To say that someone is not willing to cause (even) that much violence is to say that they would certainly not commit any act more violent than that, therefore that they would not commit any act of violence at all.

The semantic link between the original and target domain implies that every metaphor can be broken up into non-abstract elements, therefore the primary condition for being able to ground and understand metaphorical expressions is to have already grounded the necessary non-abstract symbols, which we optimistically claim may be well achievable through the methods we described earlier.

For the second step, we need to know how an artificial agent may be able to understand the translation of original domain into the target domain. In humans we can claim this happens through being exposed to idioms and such simple metaphorical expressions in their interaction with others, which, if the embodied symbol-emergence based approach we have been advocating for holds, is likely to happen in any learning artificial agents with proper grounding capabilities described at the end of section 2.2. Here, it is also worth noting that some extremely common idioms, such as the one used in our example above, may also work the other way around: an agent, human or artificial, commonly exposed to the use of this particular idiom, may, for example, learn that harming a fly is the lowest form of violence through being exposed to the metaphor.

Another approach that may yet better coincide with our requirements for embodiment and human cognition simulation is found in the works of Lakoff and Johnson (1980, 1999, 2003), namely in their notion of a conceptual metaphor. The latter argues that metaphors do not pertain only to language but to cognition in general. That is to say that humans tend to utilise metaphors not only to express themselves but also to think about things on conscious and unconscious levels. The latter concept of unconscious processing of metaphors is called functional embodiment and observes that certain concepts, including conceptual metaphors, are used automatically in cognitive processes without conscious awareness of the agent, as opposed to only being understood on an intellectual level (Lakoff 1987). This leads to some interesting implications about metaphorical mapping (i.e. the mental transition from the source domain to target domain, as well as translation from target to source) as a subconscious cognitive tool used automatically to process and describe perceptions and experience, as well as to interpret verbal inputs in metaphoric form, which may give rise to the category of metaphorical simulations, which we can fit in with Barsalou's (2008) mental simulation categories (our thanks to one of the anonymous reviewers for pointing this out). Lakoff and Johnson (1980/2003) somewhat controversially go as far as to say that metaphor mapping may be directly related to the way our brains are mapped – this, if true, practically guarantees that if proper grounding is achieved as we have described in section 2, conceptual metaphor mapping will emerge in an embodied, interactive agent.

Lastly, we may also conceive of how literary metaphors may be grounded – through exposure to common idioms, an agent learns what metaphorical meanings certain symbols commonly hold, for example that fire is often metaphorical of life, or flame of passion, etc. The process is completely analogous to one of function affordance by Glenberg and Robertson (2000) that we have described earlier. Further, there are certain metaphors in literature that are entirely unique and their meaning is speculated about by literary analysts – in these cases it is perfectly acceptable for an artificial agent to have ability of exercising such speculations, making non-stupid guesses based on its previous experience of metaphors, as we do not expect it to possess a magical insight into the mind of the metaphor's creator. However, this does not need be the case; an important part of Lakoff and Johnson's idea of conceptual metaphors is that metaphors may be grounded in simpler metaphors (equivalences, such as "love is war") that can then produce a virtual infinity of related metaphors (see also Pinker, 2007), and are themselves grounded in concrete concepts that are perceptually and experientially accessible and then serve as source domains to be related and mapped into target domains when metaphors are formed or analysed.

4. *Conclusions*

What we ultimately provided here is a theoretical approach to symbol grounding that merges compatible elements of prior prominent models of symbol grounding, including embodied agents, long-term learning that emulates human first language learning process, and symbol emergence theory, which has the explanatory power with which it can satisfy Bringsjord's (2014) non-stupidity condition.

The explanatory power lies in being exposed to a vast amount of language symbols through interaction with the environment over a long period of time, through which process an artificial agent builds a database of various contextual uses of individual symbols and from it learns to correctly determine the meaning of a symbol in certain context – a process which allows for grounding of specific contextual affordances of symbols, such as metaphoric ones, and predicting (guessing) the meaning of symbols in first-time-seen contexts.

Despite being quite successful at explaining these already high-order levels of grounding, the approach has certain limitations. For example, it remains to be determined, how certain elements of human communication, such as sarcasm, irony, or humour could be understood or grounded by artificial intelligence, even though we have hinted that the solution may lie in pre-given capabilities related to identifying tone of discourse and similar elements. Therefore we have approached satisfying the non-stupidity condition, but there are still certain questions to be answered before the explanatory power of this working model is entirely adequate.

References

Barsalou, L. W. 1999. "Perceptual symbol systems." *Behavioural Brain Science* 22: 577–660.

Barsalou, L. W. 2008. "Grounded Cognition." *Annual Review of Psychology* 59: 617–645.

Bringsjord, S. 2014. "The symbol grounding problem ... remains unsolved." *Journal of Experimental & Theoretical Artificial Intelligence* 27 (1): 63–72.

Cangelosi, A. and Riga, T. 2006. An Embodied Model for Sensorimotor Grounding and Grounding Transfer: Experiments with Epigenetic Robots." *Cognitive Science* 30: 673–689.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. 2017. "The Galilean Challenge: Architecture and Evolution of Language." *J. Phys.: Conf. Ser.* 880 012015.

Cowley, S. J. 2007. "How human infants deal with symbol grounding." *Interaction Studies* 8 (1): 83–104.

Davidsson, P. 1993. "Toward a General Solution to the Symbol Grounding Problem: Combining Machine Learning and Computer Vision." In *AAAI Fall Symposium Series, Machine Learning in Computer Vision: What, Why and How?*, 157–161.

Fodor, J. A. 1980. "Methodological solipsism considered as a research strategy in cognitive psychology." *Behavioral and Brain Sciences* 3: 63–69.

Glenberg, A. M. in Robertson, D. A. 2000. "Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning." *Journal of Memory and Language* 43: 379–401.

Groefsema, M. 1995. "Can, may, must and should: A relevance theoretic account." *Journal of Linguistics* 31: 53–79.

Guazzini, J. 2017. "An Epistemological Approach to the Symbol Grounding Problem." In V. C. Müller (ed.). *Philosophy and Theory of Artificial Intelligence.* Leeds: Springer, 36–39.

Harnad, S. 1990. "The symbol grounding problem." *Physica D* 42: 335–346.

Hutto, D. D. and Myin, E. 2013. *Radicalizing Enactivism: Basic Minds without Content.* Cambridge: MIT Press.

Jary, M. 2010. *Assertion.* London: Palgrave Macmillan.

Jary, M. in Kissine, M. 2014. *Imperatives.* Cambridge: Cambridge University Press.

Kissine, M. 2013. *From Utterances to Speech Acts.* Cambridge: Cambridge University Press.

Lakoff, G. 1987. *Women, Fire, and Dangerous Things.* Chicago: The University of Chicago Press.

Lakoff, G. and Johnson, M. 1980/2003. *Metaphors We Live By. With Afterword 2003.* Chicago: The University of Chicago Press.

Lakoff, G. and Johnson, M. 1999. *The Embodied Mind and Its Challenge to the Western Thought.* New York: Basic Books.

Maturana, H. R. and Varela, F. J. 1992. *The tree of knowledge: the biological roots of human understanding.* Boulder: Shambhala Publications.

Mayo, M. 2003. "Symbol Grounding and its Implication for Artificial Intelligence." *Twenty-Sixth Australian Computer Science Conference*, 55–60.

Müller, V. C. 2015. "Which Grounding Problem Should We Try to Solve?" *Journal of Experimental & Theoretical Artificial Intelligence* 27 (1): 73–78.

Noe, A. 2004. *Action and Perception.* Cambridge: MIT Press.

Papafragou, A. 1998. *Modality and the Semantics-Pragmatics Interface.* London: University College London.

Pezzulo, G. et al. 2013. "Computational Grounded Cognition: a new alliance between grounded cognition and computational modelling." *Frontiers in Psychology* 3: 1–11.

Pierce, B. 2017. "How Are Robots' Reasons for Action Grounded?" In V. C. Müller (ed.). *Philosophy and Theory of Artificial Intelligence.* Leeds: Springer, 73–80.

Pinker, S. 2007. *The Stuff of Thought.* London: Penguin Publishing Group.

Quine, W. V. O. 1960. *Word and Object.* Cambridge: MIT Press.

Regier, T. 1992. *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization.* Berkeley: Department of Computer Science, University of California at Berkeley.

Rodriguez, D. et al. 2011. "Meaning in Artificial Agents: The Symbol Grounding Problem Revisited." *Minds and Machines* 22 (1): 25–34.

Searle, J. R. 1980. "Minds, brains and programs." *Behavioral and Brain Sciences* 3: 417–457.

Shepard, R. N. and Cooper, L. A. 1982. *Mental images and their transformations.* Cambridge: MIT Press/Bradford.

Steels, L. 2008. "The symbol grounding problem has been solved, so what's next?" In M. de Vega. et al. (eds.). *Symbols and embodiment: Debates on meaning and cognition.* Oxford: Oxford University Press, 223–244.

Steels, L. in Vogt, P. 1997. "Grounding adaptive language games in robotic agents." In C. Husbands and I. Harvey (eds.). *Proceedings of the 4th European Conference on Artificial Life.* Cambridge: MIT Press.

Šetar, N. 2020a. *A Monosemic Account of Modality in Speech Act Theory. MA Thesis.* Maribor: Univerza v Mariboru.

Šetar, N. 2020b. *Utemeljevanje simbolov in pragmatika v umetni inteligenci. MA Thesis.* Maribor: Univerza v Mariboru.

Taddeo, M. in Floridi, L. 2005. "Solving the symbol grounding problem: A critical review of fifteen years of research." *Journal of Experimental and Theoretical Artificial Intelligence* 17 (4): 419–445.

Taddeo, M. in Floridi, L. 2007. "A Praxical Solution of the Symbol Grounding Problem." *Minds & Machines* 17: 369–389.

Tangiuchi, T. et al. 2016. "Symbol emergence in robotics: a survey." *Advanced Robotics* 30 (11–12): 706–728.

Tangiuchi, T. et al. 2019. "Survey on frontiers of language and robotics." *Advanced Robotics* 33 (6): 2–31.

Varela, F. J., Thompson, E. and Rosch, E. 1991. *The Embodied Mind: Cognitive Science and Human Experience.* Cambridge: The MIT Press.

Vogt, P. 2007. "Language Evolution and Robotics, Issues on Symbol Grounding and Language Acquisition." In A. Loula et al. (eds.). *Artificial Cognition Systems.* Hershey: Idea Group Publishing, 176–209.

Ziemke, T. 1999. "Rethinking Grounding." In A. Riegler et al. (eds.). *Understanding Representation in the Cognitive Sciences.* New York: Plenum Press, 177–180.

Croatian Journal of Philosophy
Vol. XXII, No. 64, 2022
https://doi.org/10.52685/cjp.22.64.7
Received: September 1, 2021
Accepted: March 31, 2022

Strict Conditionals. Replies to Lowe and Tsai

JAN HEYLEN
KU Leuven, Leuven, Belgium
LEON HORSTEN
University of Konstanz, Konstanz, Germany

Both Lowe and Tsai have presented their own versions of the theory that both indicative and subjunctive conditionals are strict conditionals. We critically discuss both versions and we find each version wanting.

Keywords: Strict conditionals; indicative conditionals; subjunctive conditionals.

1. *Introduction*

In the vast literature on conditionals there are some theories that give a unified account of both indicative and subjunctive conditionals in natural language — see Bennett (2003: ch. 23) for a discussion of the unified accounts Davis (1979, 1983), Stalnaker (1975, 1984), Ellis (1978, 1984) and Edgington (1995, 2003). Lowe (1983, 1995) and Tsai (2016) have both also proposed a unified theory of conditionals.[1] Whereas Ellis and Stalnaker favour a theory according to which conditionals are 'variably strict conditionals,' which are of the form $\phi \mathrel{\Box\!\!\rightarrow} \psi$ and which are given a similarity semantics (Stalnaker 1968; Lewis, 1973), Lowe and Tsai favour a theory according to which conditionals are 'strict conditionals', which are of the form $\phi \prec \psi$ or, equivalently, $\Box(\neg\phi \vee \psi)$ (Lewis 1912). Likewise, Daniels and Freeman (1980), Warmbrod (1983), von Fintel (2001) and Gillies (2007) also prefer the analysis in terms of strict con-

[1] Lowe (1979, 1980) defends a unified theory based on the claim that, for counterfactuals belonging to so-called 'Adams pairs' (Adams 1970) one can find a future-tense, indicative conditional that is equivalent to it.

ditionals, although they limited their analyses to certain classes of conditionals, namely subjunctive or epistemic conditionals. The discussion between proponents of the strict conditionals analysis and the variably strict conditionals analysis is set against the background of fundamental discussion about the respective roles of semantics and pragmatics. In this short critical reply we will focus only on the work of Lowe and Tsai.

Since both Lowe and Tsai defend that natural language conditionals are a kind of strict conditionals, they have to deal with the so-called 'paradoxes of strict implication,' which can best be understood against the background of the so-called 'paradoxes of material implication.' Suppose that natural language conditionals of the syntactical form 'If *p*, then *q*' are material conditionals or 'implications,' i.e., they are of the logical form $p \supset q$, which is equivalent to:

(1) $$\neg p \vee q$$

Lewis (1912: 524) noted that it would follow that one has to accept 'If Caesar did not die, then the moon is made of green cheese.' He also noted that it would follow that one has to accept 'If the moon is not made of green cheese, then Caesar died' (Lewis 1912: 527). Lewis (1912: 529) generalizes this by pointing out that the following implications hold:

(2) $$\neg p \supset (p \supset q)$$

(3) $$q \supset (p \supset q)$$

The above are known as paradoxes of material 'implication.' Lewis contrasted the material conditionals $p \supset q$ with strict conditionals or 'implications' $p \prec q$ and, correspondingly, extensional disjunctions of the form (1) with intensional disjunctions of the following form:

(4) $$\Box (\neg p \vee q)$$

It can be proved that, if one replaces by in (2)–(3), then the resulting formulas are no longer valid. However, Lewis and Langford (1932, 174) noted that there are also paradoxes of strict 'implication,' namely:[2]

(5) $$\neg \Diamond p \prec (p \prec q)$$

(6) $$\Box q \prec (p \prec q)$$

We are now ready to turn to Lowe's and Tsai's respective theories.

2. *Lowe on conditionals*

Lowe wants to have a theory according to which all natural language conditionals are a kind of strict conditionals.[3] However, he does accept that (5) presents a problem. Lowe (1995, 48) offers the following ex-

[2] While Lewis and Langford (1932) prove this for their so-called 'non-normal' system of modal logic, they are also theorems in the weakest system of 'normal' modal logic.

[3] Lowe (1995: 50) does rule out so-called 'Dutchman conditionals', e.g., 'if that is a Ming vase, then I am a Dutchman.'

ample: 'If I had bought a ticket, I would have won,' where it is assumed that it is impossible for me to buy a ticket. Another example is the following: 'If 0 = 1, then the sun will shine tomorrow' (Heylen and Horsten 2006: 538). As a putative solution, Lowe (1995: 48) considers the following variation:

(7) $$\Box (\neg p \vee q) \wedge \Diamond p$$

Lowe (1995: 48) offers the following counterexample to (7):

(8) If *n* were the greatest natural number,
then there would be a natural number greater than *n*.

However, as Heylen and Horsten (2006: 539) observe, *n* is a free variable. This means that we cannot directly talk about the truth or falsity of the example, but only indirectly, namely via the truth or falsity of its universal closure. Better examples in this respect are the following: 'If 0 = 1 and 1 = 1, then 1 = 1' and 'If Frege Arithmetic is consistent, then Peano Arithmetic is consistent' (Heylen and Horsten 2006: 542). In both cases there is a logical connection between the antecedent and the consequent, namely an application of the rule of conjunction elimination and (a corollary of) Frege's Theorem respectively. Lowe (1995: 49) revised his solution by putting forward the following second and final variation:

(9) $$\Box (\neg p \vee q) \wedge (\Diamond p \vee \Box q)$$

While according to Lowe natural language conditionals all have (9) as their logical form, Lowe (1995: 49-51) states that the interpretations of the modal operators $\Box$ and $\Diamond$ in (9) can vary. Lowe notes that the modal operators can be given the redundant interpretation, which turns (9) into $(\neg p \vee q) \wedge (p \vee q)$ or, equivalently, *q*. As an example, he points to so-called 'biscuit conditionals', e.g. 'there are biscuits on the sideboard, if you want some.' Paradigm examples of (present-tense) indicative conditionals involve an epistemic reading of the modal operators, whereas paradigm examples of counterfactuals involve an alethic reading of those modal operators. An epistemic reading of the $\Box$ operator is 'it is certain that.' An alethic reading of the $\Box$ operator is 'it is inevitable that.'

Lowe (1995) has been criticized by Heylen and Horsten (2006: 539), who provide the following counterexample:[4]

(10) If 2 = 3, then 2 + 1 = 3 + 1.

As before, there is a logical connection between the antecedent and the consequent, namely an application of Leibniz's law and the law of self-identity. Furthermore, Heylen and Horsten (2006: 540-545) claim that there is no propositional condition *X* that can be expressed in

[4] Heylen and Horsten (2006: 539) also gave the following counterexample: 'If I am my father, then my father is my father's father.' However, as Lowe (2008: 529) pointed out: according to Macbeath (1982) it is possible that someone is his own father in some time travel scenario's.

terms of proposition letters p, q, the modal operator $\Box$, and the classical propositional connectives such that $\Box(\neg p \vee q) \wedge X$ is *exactly* strong enough in the sense that there are no intuitively false conditionals that have that logical form and that there are intuitively true conditionals that lack that logical form. They worked with modal system **S5** in the background, because in **S5** every formula is provably equivalent to a 'flat' formula, which does not contain modal operators inside the scope of other modal operators. The latter is important, because they claim that 'it would be scarcely imaginable that the correct interpretation of conditionals essentially involves nested modalities' (Heylen and Horsten 2006: 540). Against this, Tsai (2016) claims that a proper unified theory of conditionals involves irreducibly nested modalities. We will return to this in section 3.

In reply, Lowe (2008) formulated a methodological criticism and a substantive criticism. Let us take these in turn.

The methodological criticism was twofold: first, he accused Heylen and Horsten to rely on a very narrow selection of examples and, second, he claimed that their examples are not conditionals that are ordinarily used in everyday conversation (Lowe 2008: 528). But a *tu quoque* response can be given. After all, Lowe (1995: 48) gave only one example against the hypothesis that the logical form of conditionals is captured by (7), and this is a 'mathematical truism.' A second response is that the use of conditionals like (10) are more wide-spread in 'mathematical English.' There are plenty of examples of conditionals with impossible antecedents and consequents in a textbook on computability and (meta-)logic (Boolos et al. 2007: 38, 40, 97, 126, 132, 134, 154, 160, 192, 223, 227, 228, 271, 284, 303) and in a textbook on algebra (Givant and Halmos 2009: 12, 215, 336, 474). For instance, Givant and Halmos (2009: 215) write the following:

> If q were a strictly smaller upper bound of E in B, then $p - q$ would be a non-zero element of B, and therefore above a non-zero element r of A, by density.

However, it is clear that very often or almost always those kind of conditionals are used in the following type of reasoning: 'If ϕ were the case, then ψ would be the case. But ψ is not the case. Therefore, ϕ is not the case.' This suggest that the following variation on (10) would have been more in accordance with the above mathematical practice:

(11) If $2 = 3$, then $2 - 1 = 3 - 1$.

It is easy to see how such a counterpossible can figure in a proof that leads to $0 = 1$, contradicting an axiom of arithmetic and, hence, leading to the conclusion that $2 \neq 3$. A third response begins by admitting that 'mathematical English' is not colloquial English, although we think that it is a fundamental mistake to draw a sharp distinction between the two. Moreover, dialectically speaking, one is forced to go look for examples from logic or mathematics or metaphysics. Otherwise, Lowe could have claimed that the antecedent is not impossible on a narrow sense of impossibility or that the consequent is not possibly false on a narrow sense of possibility. For instance, 'If I had participated as an athlete in the Olympic Games, I would first have passed the Olympic Trials.' It is open for an objector to claim that the antecedent is not metaphysically impossible.

The substantive criticism starts from the observation that the use of '=' obscures whether one is dealing with an indicative ('is equal to') or a subjunctive ('were equal to').

Suppose that the conditional is in indicative mode: 'If 2 is equal to 3, then 2 + 1 is equal to '. Then the modal operators have to be interpreted as epistemic operators. Lowe (2008: 529–530) suggests the following reading: '□' means 'it is certain that' and '◊' means 'it is uncertain that not'. Furthermore, Lowe distinguishes between *real* (un)certainty and *feigned* (un)certainty. While there is real certainty that $2 \neq 3$, Lowe suggests that in some context uncertainty that $2 \neq 3$ may be feigned. In those contexts his theory predicts that 'If 2 is equal to 3, then 2 + 1 is equal to 3 + 1' is acceptable after all.

Suppose that the conditional is in subjunctive mode: 'If 2 were equal to 3, then 2 + 1 were equal to 3 + 1'. Lowe considers a possible world w in which only the numbers **0**, **1** and **2** exist. Moreover, in w '3' refers to **2**, so the antecedent of the conditional is true. Furthermore, in w the adding-one function is partial: only if the input is **0** or **1** is the output defined (and it the standard outcome). In this world there would be no number corresponding to '2+1'. Lowe claims that the consequent of the conditional is therefore false in that world.[5] In addition, Lowe (2008, 530) claims that a similar strategy works can be used to show that there is a possible world in which (11) is false.

These last considerations by Lowe lead to radical views in ontology and semantics. One implicit assumption is that natural numbers exist only contingently. Most platonists are not happy with that assumption. Another implicit assumption is that the natural numbers have possible existence independently of other natural numbers. Structuralists disagree with this assumption. So, nominalist structuralism is not an option here. Lowe is also assuming that not all mathematical terms are rigid designators (i.e. terms that designate the same object in all possible worlds in which that object exists and that never designate any other object), while Kripke (1980) illustrated the notion of a rigid designator with the help of arithmetical terms (e.g. 'the smallest prime').

Finally, it appears to have escaped Lowe's notice that his special possible world would also make (8) false. But Lowe had claimed that it is intuitively true. Moreover, he has used the intuitive truth of (8) to argue against (7).

In conclusion, Lowe's version of the theory that natural language conditionals are strict conditionals fails to convince.

[5] This assumes that an atomic sentence is false at a world if at least one the terms occurring in it does not denote anything in that world.

3. *Tsai on strict conditionals*

Tsai (2016: 78) starts with (7), which he labels 'Default'. He defends the extra condition $\Diamond p$ by reference to the so-called Ramsey Test (Ramsey 1929: 247): $\Diamond p$ expresses that p is an epistemic possibility, so it is open to add it to the 'stock of knowledge'. Details aside, what matters most here is that the modal operators in (7) are given an epistemic reading. We should also mention that Tsai gives a formal interpretation of the modal language that does not make use of Kripke models but rather of models in the style of Becker (1952), which Tsai has further developed in earlier work (Tsai 2012). However, we will not go into the details, because Tsai (2012: 107, 112) points out that there is an 'isomorphism' between Beckerian 'hi-worlds' and a defined 'sub-hi-world' relation and Kripkean frames, which contain worlds and an accessibility relation between them.[6]

As we have seen in section 2, Lowe offered (8) as a counterexample to (7). Tsai (2016: 82) agrees that (8) is intuitively true. The solution of Lowe was to accept (9). Tsai (2016: 79) observes that a consequence of (9) is that one has to accept one of the paradoxes of strict implication, namely (6). On this basis, Tsai rejects Lowe's theory and proposes his own solution.

Tsai (2016: 80) proposes what he labels 'Unified':

(12) $(\neg p \vee q)$ or $(\Box(\neg p \vee q) \wedge \Diamond p)$ or $(\Box\Box(\neg p \vee q) \wedge \neg\Diamond p \wedge \Diamond\Diamond p)$

The idea is that the logical form of a given natural language conditional is one of the disjuncts of (12). Like Lowe, he accepts that sometimes a natural language conditional can have the logical form of a material implication, and he also agrees that this is a rare case. So, it is mainly about the last two disjuncts. The implicit assumption here is that modal principle 4, namely $\Diamond\Diamond\phi \rightarrow \Diamond\phi$, is not valid, because otherwise the third disjunct would be contradictory. This means that according to Tsai the modalities involved are irreducibly nested, contrary to the assumption made by Heylen and Horsten (2006). This also entails that the modal operators in (12) cannot be understood as expressing logical modalities (Burgess 1999), mathematical modalities (Hamkins and Linnebo 2019) or metaphysical modalities (Williamson 2016), which figured prominently in the discussion of Lowe, since adequate systems for those notions all contain at least modal principle 4. Next, Tsai (2012: 80-81) makes a puzzling claim, namely that, if one takes material implica-

[6] Tsai's claim needs to be qualified slightly: to each Beckerian model there corresponds a Kripkean frame with a *serial* accessibility relation, whereas there are Kripkean frames with a non-serial accessibility relation (i.e. at least one of the worlds does not have access to any world). The reason is that Tsai (2012: 109) stipulates that a hi-world is of the form $\langle U^0, U^1, \ldots\rangle$, where U^0 is an element of the domain D of the model and, for each $i \geq 1$, U^i is an element of $(\mathcal{P}^*)^i\,(D)$, with $\mathcal{P}^*(X) = \mathcal{P}(X) \setminus \emptyset$. Given Tsai's epistemic reading of the modal operator, this qualification is not important, since it is generally accepted that non-serial accessibility relations are inadequate for modelling rational belief and knowledge.

tions out of consideration and if one accepts the validity of $\Box\phi \rightarrow \phi$ or, equivalently, $\phi \rightarrow \Diamond\phi$, (12) can be 'reduced' to what he labels 'Core':

(13) $(\Box(\neg p \vee q) \wedge \Diamond p)$ or $\wedge$ $(\Box\Box\ (\neg p \vee q) \wedge \Diamond\Diamond p)$

However, while (13) logically follows from the last two disjuncts of (12), the converse is not true, even on the assumption that Tsai mentions. In any case, Tsai (2016: 81) labels the second disjunct of (13) 'Subjunctive' and he adds that it is relevant when '*p* is deemed impossible.' Given that $\Diamond$ is supposed to be the epistemic possibility operator, $\Diamond\Diamond$ has to be understood as the *epistemic* possibility of the epistemic possibility.

With the above theory Tsai (2016: 82) tries to account for the intuitive truth of (8). He invites us to imagine a 'pseudo-mathematical system' in which there is a greatest natural number *n*. This already raises two questions. First, how is imagination related to the epistemic possibility of the epistemic possibility? For a mathematician who has *reflective* knowledge about there not being the largest natural number there is no epistemic possibility that there is the epistemic possibility of *n* being the largest natural number. Second, what *are* pseudo-mathematical systems? Perhaps Tsai could take a cue from Lowe and imagine a world in which not all numbers exist and/or in which mathematical vocabulary is interpreted in a non-standard way. But even if these questions can be answered satisfactorily, there is the problem that there is no Beckerian or Kripkean model in which (i) the antecedent of (8) is possibly possible and (ii) (8) is necessarily necessary. The reason is that the consequent, which can be formalized as $\exists x x > n$, is logically equivalent to the negation of the antecedent, which can be formalized as $\neg\exists x x > n$. Suppose now that there is some possibly possible world at which the antecedent is true.[7] Then by the necessity of the necessity of the material implication the consequent also has to be true at that possibly possible world. Yet, there is no Beckerian or Kripkean world in which logical contradictions are true, even with the countenance of the ontological and semantical views Lowe was willing to resort to. Therefore, (8) has to be false on Tsai's theory.

For another counterexample to Tsai's theory, consider first the following conditional:

(14) If 1 = 1, then 1 = 1.

calls it a 'truism' and he uses it to argue against Hitchcock (1998: 25), to whom he attributes the view that the logical form of a conditional is the following:[8]

(15) $\Box\ (\neg p \vee q) \wedge (\Diamond p \wedge \Diamond\neg\ q)$

[7] By 'possibly possible' we mean that it belongs to U^2 (Becker) or that it is accessible from some accessible world (Kripke).

[8] Note that Hitchcock (1998) is really talking about *logical consequence*. He adds the condition that it is possible that the *premises* are true and the condition that it is possible that the *conclusion* is false.

But if Tsai is willing to accept (14), then he should also be willing to accept the following:

(16) If $1 \neq 1$, then $1 \neq 1$.

Surely, (16) is no less a truism. Note that one does not even need the controversial rule of contraposition but only the observation that the antecedent and the consequent are the same. Yet, there is no Beckerian or Kripkean world in which logical contradictions are true. By the way, we take (16) also to be a counterexample to Lowe's theory of subjunctive conditionals.

Neither Lowe's nor Tsai's version of the theory that natural language conditionals are strict conditionals has withstood critical scrutiny.

References

Adams, E. W. 1970. "Subjunctive and indicative conditionals." *Foundations of Language* 6 (1): 89–94.

Becker, O. 1952. *Untersuchungen Über den Modalkalkül*. Westkulturverlag A. Hain.

Bennett, J. 2003. *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Boolos, G., J. Burgess, R. P., and C. Jeffrey. 2007. *Computability and Logic*. Cambridge: Cambridge University Press.

Burgess, J. P. 1999. "Which modal logic is the right one?" *Notre Dame Journal of Formal Logic* 40 (1): 81–93.

Daniels, C. B. and J. B. Freeman. 1980. "An analysis of the subjunctive conditional." *Notre Dame Journal of Formal Logic* 21 (4): 639–655.

Davis, W. A. 1979. "Indicative and subjunctive conditionals." *Philosophical Review* 88 (4): 544–564.

Davis, W. A. 1983. "Weak and strong conditionals." *Pacific Philosophical Quarterly* 64 (1): 57.

Edgington, D. 1995. "On conditionals." *Mind* 104 (414): 235–329.

Edgington, D. 2003. "What if? Questions about conditionals." *Mind and Language* 18 (4): 380–401.

Ellis, B. 1978. "A unified theory of conditionals." *Journal of Philosophical Logic* 7 (1): 107–124.

Ellis, B. 1984. "Two theories of indicative conditionals." *Australasian Journal of Philosophy* 62 (1): 50–66.

Gillies, A. S. 2007. "Counterfactual scorekeeping." *Linguistics and Philosophy* 30 (3): 329–360.

Givant, S. and P. Halmos. 2009. *Introduction to Boolean Algebras*. Dordrecht: Springer.

Hamkins, J. D. and O. Linnebo. 2019. "The modal logic of set-theoretic potentialism and the potentialist maximality principles." *Review of Symbolic Logic* 1–36.

Heylen, J. and L. Horsten. 2006. "Strict conditionals: A negative result." *Philosophical Quarterly* 56 (225): 536–549.

Hitchcock, D. 1998. "Does the traditional treatment of enthymemes rest on a mistake?" *Argumentation* 12 (1), 15–37.

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

Lewis, C. I. 1912. "Implication and the algebra of logic." *Mind* 21 (84): 522–531.

Lewis, C. I. and C. H. Langford. 1932. *Symbolic Logic*. New York: Century Company.

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

Lowe, E. 1980. "Reply to Davis." *Analysis* 40 (4): 187–190.

Lowe, E. J. 1979. "Indicative and counterfactual conditionals." *Analysis* 39 (3): 139–141.

Lowe, E. J. 1983. "A simplification of the logic of conditionals." *Notre Dame Journal of Formal Logic* 24 (3): 357–366.

Lowe, E. J. 1995. "The truth about counterfactuals." *Philosophical Quarterly* 45 (178): 41–59.

Lowe, E. J. 2008. "'If 2 = 3, then 2 + 1 = 3 + 1': Reply to Heylen and Horsten." *Philosophical Quarterly* 58 (232): 528–531.

Macbeath, M. 1982. "Who was Dr Who's father?" *Synthese* 51 (3): 397–430.

Ramsey, F. P. 1929. "General propositions and causality." In F. P. Ramsey (ed.). *The Foundations of Mathematics and other Logical Essays*. Kegan Paul, Trench, Trübner, 237–255.

Stalnaker, R. 1975. "Indicative conditionals." *Philosophia* 5 (3): 269–286.

Stalnaker, R. 1984. *Inquiry*. Cambridge: Cambridge University Press.

Stalnaker, R. C. 1968. "A theory of conditionals." In N. Rescher (ed.). *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)*. Oxford: Blackwell, 98–112.

Tsai, C. 2012. "The genesis of hi-worlds: Towards a principle-based possible world semantics." *Erkenntnis* 76 (1): 101–114.

Tsai, C. 2016. "Becker, Ramsey, and hi-world semantics. Toward a unified account of conditionals." *Croatian Journal of Philosophy* 16 (1): 69–89.

von Fintel, K. 2001. "Counterfactuals in a dynamic context." In M. Kenstowicz (ed.). *Ken Hale: A life in language*. Cambridge: MIT Press, 123–152.

Warmbrod, K. 1983. "Epistemic conditionals." *Pacific Philosophical Quarterly* 64 (3): 249–265.

Williamson, T. 2016. "Modal science." *Canadian Journal of Philosophy* 46 (4-5): 453–492.

Croatian Journal of Philosophy
Vol. XXII, No. 64, 2022

Book Review

John Perry, Frege's Detour: An Essay on Meaning, Reference, and Truth, *Oxford University Press, 2019, xii + 148 pp.*

In 1872 Frege wrote an essay titled "On sense and reference" where he presented his sense and reference theory of meaning. Since then, the essay has gained a canonical status in the philosophy of language literature, and philosophy students all over the world have the essay as a reading assignment in the philosophy of language classes. The noted philosopher of language John Perry does not share this sentiment. On the contrary, he thinks that "this essay put philosophy on detour" (1). In the ten chapters that this book consists of, Perry explains what that detour is and gives his solution to how we can get back on track while simultaneously keep what Frege got right about meaning.

The first chapter is introductory. There Perry lays out Frege's detour. It was the doctrine of indirect reference, his solution to a difficulty for his sense and reference theory. The difficulty is created by indirect discourse and attitude report sentences where the principle that the reference of a complex expression like a sentence is determined by the reference of its parts does not seem to hold. A corollary of the principle is that a part of a complex expression can be replaced by another one that is co-referring without affecting the reference of the complex expression. Indirect discourse and attitude reports, however, do not permit that. To use Perry's example, the sentence "Smith believes that Berkeley is west of Santa Cruz" according to the principle and its corollary, keeps its reference, the truth value True, when the embedded part that stands for Smith's true belief is replaced by another true sentence, that Mogadishu is the main capital of Somalia, despite Smith not believing in this. Frege's solution is that sentences when embedded in an indirect discourse or an attitude reports do not refer to their truth value but they refer either to what they quote or their usual sense, the Thought they express. So the substitution is not permitted in such sentences while the compositionality principle is preserved. Perry rejects the doctrine of indirect reference because it did not, contrary to Frege, give a solution and because it has helped to spread and legitimize two thesis about truth and cognition that are in Perry's opinion false: (A) that there is a unique proposition that captures the sentence's content, its truth-conditions, which carries its cognitive significance in the sense that it is what the speaker of the sentence means and believes and it is the reference of embedded sentences in indirect discourse and attitude report, and that (B) attitudes such as beliefs are a relation between an agent and a proposition. An alternative way, says Perry, that can help us stay from the detour and the faulty assumptions is found in Frege's earlier major work *Begriffsschrift* where he had a different theory of meaning that he later abandoned for the sense and reference theory.

In the second chapter, Perry lays out the semantic theory in *Begriffschrift* that Frege had abandoned for the sense and reference theory. The theory of conceptual content was the theory of meaning under which Frege operated while writing *Begriffschrift*. It acted as the semantic framework within which Frege developed first and second-order logic. Perry highlights that it was largely implicit, so what he says is his interpretation of Frege's ideas in *Begriffsschrift*. According to the theory, as it names says, what language expressions refer to is conceptual content. The conceptual content of a sentence is circumstance (*Umstand*). It possesses truth-value and if true is also a fact. Perry tells us that Frege never elaborates in *Begriffschrift* what circumstances are. He just states several times that sentences refer to them. Here Perry goes into interpretive mode. He attributes to Frege the view of circumstances as potential facts and complexes made up of objects, properties, and relations that objects have either with other objects or properties. Perry justifies this reading of Frege's circumstances by explaining that non-idealist philosophers in the 19th century took a realist stance of relations and designated them as the third component, next to objects and properties, that make up a fact. Frege here also held to the compositionality principle. The conceptual content of an expression is determined by the conceptual content of its parts. He bifurcates sentences into names and predicates. The conceptual content of names are objects and of predicates properties. Another crucial aspect of the theory that Perry mentions is that sentences with the same conceptual content have the same logical consequences.

The third chapter Perry devotes to the reason why Frege rejected the conceptual content theory and which led him to develop his more famous theory, the reason being that he concluded that circumstances do not provide the truth-conditions of sentences which carry their cognitive significance. What led Frege to this conclusion, explains Perry, is the general issue of identity that his *Begriffschrift* theory was unable to solve. Frege's dealings with identity started out with two identity problems that were implicitly in the background of Section 8 of *Begriffschrift* and culminated in a general identity problem found in his later article "Concept and function". Perry gives a detailed account of the identity problem and Frege's solution to them. For good measure he adds an identity statement problem formulated by the philosopher George Wilson. The two identity problems in *Beggriffschrift*, which Perry dubs the Name problem and the Co-instatiation problem, are about identity statements between names. The identity statements with the same circumstance, "Hesperus = Hesperus" and "Hesperus = Phosphorus", must have the same logical consequence but they do not. The first one is trivial, the second informative, and from the second one can infer that Hesperus and Phosphorus refer to the same thing. This is the Name problem. When an additional premise is added to those sentences, e.g. that the reference of "Hesperus" is determined by pointing to the first planet that appears in the evening sky and saying, "That is Hesperus" and that the reference of "Phosphorus" is determined by pointing out the last planet that appears in the morning sky and saying "That is Phosphorus", the same information must be inferred as they have the same logical consequence, but it is not. The second sentence and the additional premise together entail that the first planet that appears on the sky and the last planet to disappear from the morning sky are the same, but not the first one. This is the Co-instatiation problem. In Section 8, Frege, next to identity, a relation between objects, introduces a new kind of identity relation that he calls the identity of content, which is a relation between names that have the same conceptual content. To distinguish it symbolically from identity, he uses the ≡ symbol to represent it. Perry notes that this is the only place in *Beggriffschrift* that the distinction and the ≡ symbol appear. He interprets the introduction of this distinction and the writing of Section 8 as only making sense if Frege had the Name and the Co-instatiation problem at the back of his mind. The solution is that the identity statements are actual identity of content statements, "Hesperus ≡ Hesperus" and "Hesperus ≡ Phosphorus". Since they have different contents, they have different logical consequences. Here is where it becomes problematic for the conceptual content theory. The Wilson problem is the problem of reflexive relations other than identity, e.g. if we know there is a planet "Hesperus", we can infer that "Hesperus is the same size as Hesperus" is a true sentence, but without more information, we are unable to know that "Hesperus is the same size as Phosphorus". Frege's solution cannot solve this problem. Neither can it be used to solve the General problem of identity, which is that sentences that refer to the same circumstances do not have the same consequences although they should if circumstances are their conceptual content. This problem finally convinced Frege to give up on circumstances and the conceptual content theory. Perry thinks that the rejection was premature.

In the forth chapter, Perry talks about the sense and reference theory as it was presented in a series of articles written during the 1890s and the accompanying problems. In contrast to the conceptual content theory, in the sense and reference theory, reference is now done indirectly through senses who pinpoint the referent. They perform the function of carrying cognitive significance of expression that objects, predicates, and circumstances failed in Frege's earlier theory. The sense of a proper name is a property of the object it refers to. Names contribute with their senses to the sense of a sentence, a Thought, which gives its truth-condition that tells if it is true or false. Perry says that there is a continuity between the two theories, for senses are property structures with better articulated descriptions. The 1890s works shows that Frege had a sense for predicates, but he never said explicitly what it is. Perry, on the basis of Frege's later works, suggests that the sense of predicates is similar to the sense of names. It is the detailed description of the property it refers to. Perry also derives the consequence that a Thought has two existential quantifiers, one that affirms there is a unique object and one that affirms a unique property. A more problematic part of his theory are concepts and extensions, which even baffles experts on Frege. In Frege's time extension was an intuitive concept with no clear definition.

He considered them to be a special case of what he calls course of values (*Werthverläufe*). What Perry makes of it is that a course of value is a set of arguments and values determined by a function, so extensions are courses of values for concepts. They are a set of arguments and values with values being True or False. So concept is to be understood as an unsaturated function, and the extension is what turns it into a saturated function, and only those concepts that are extensionally individuated can be a reference of predicates, i.e they are properties. The problems for the sense and reference theory are the Regress problem, the problem that emerges because since Thoughts do not have objects, neither must the senses of names, but because sense of names often have then, a regress emerges finding a sense of a name not containing objects, Kerry's problem, the problem of names, who refer to objects, saturated entities, referring to properties, which are unsaturated entities, and the problem of accommodating properties that share the same extension. Some of these problem Perry tackles in eight chapter.

In the fifth chapter, Perry takes under the loop Frege's sense and reference theory how he presents it in the article "On Sense and Reference". Senses give the necessary and sufficient conditions that an object must fulfill for it to be the reference of an expression, but where commentators get it wrong according to Perry is identifying senses with modes of representation. They are a part of sense but not identical to them. He characterizes modes as functions. Their arguments are presenters, and their values are presented objects. Sense contains modes and the sense of presenters but not objects. Another thing that commentators assume is true is that Frege treated proper names as hidden descriptions, when there is no evidence for this. Frege actually tells very little about the senses of proper names, but where he does mention something what is crucial is his distinction between a perfect language that is used for scientific research where only one sense is attached to an expression and imperfect languages that are used for everyday communication where an expression has multiple senses. The purpose of sense and reference theory is to give an account of the perfect language. Frege then applies this theory also to imperfect languages whose deficiencies are tolerable in a nonscientific discourse because successful communication is possible despite of them. Here Perry says there is a place for circumstance in a semantic theory. People successfully communicate and exchange information about a thing they attach different Thoughts to because they agree about the circumstance. Thoughts exemplify truth conditions and cognitive significance, but they are poor carriers of information. This also gives a good reason why circumstances are a good candidate for being the reference of sentences, but Frege does not go in this direction. He designates truth values as the things that sentences refer to, but he does not give a good reason for this.

In the sixth chapter, Perry shows how Frege's conceptual content theory from *Beggriffschrift* can solve the identity problems that he presented in chapter two. What prevented Frege from realizing it was his adherence to the doctrine of unique content, though he came near it in Section 8 of *Beggrrifschrift* where he introduces the distinction between identity and identity of the content. The basic idea is that expressions not only convey information about the things they stand for but also information about themselves, which is often the point of using them. It also shows that the doctrine of unique content is false, for it means that an expression's truth-conditions about the objects it refers to, the usual content it conveys, are not the only truth conditions a sentence has. Perry distinguishes three truth-conditions found in *Beggriffschrift*: (1) reflexive truth-conditions under which a sentences is true, e.g. the sentence "Bratman is taller than Lawlor." is true iff there are objects *x* and *y* and a relation Ψ such that *x* and *y* are the objects to which "Bratman" and "Lawlor" refer and Ψ is the relation to which "is taller than" refers and that the circumstance that *x* has Ψ to *y* is a fact, (2) referential truth-conditions that specify how the sentence could satisfy the reflexive truth-conditions; the referential truth-condition of "Bratman is taller than Lawlor." is that the circumstance that Bratman is taller than Lawlor is a fact, and (3) hybrid truth-conditions, the conditions for some expressions that make up the sentence. Perry uses this Reflexive-referential theory as he calls it to solve the identity problems. In the Name problem, the identity sentences "Hesperus = Hesperus" and "Hesperus = Phosphorus" have the same referential truth-condition, namely that the circumstance that Venus is identical to Venus is a fact, but differ in their reflexive and hybrid truth-conditions. The reflexive and hybrid truth-conditions of "Hesperus = Phosphorus" proscribe the existence of two objects, *x* and *y*, to which names Hesperus and Phosphorus refer, while the reflexive and hybrid truth-conditions of "Hesperus = Hesperus" proscribes the existence of two objects, *x* and *y*, to which the name Hesperus refers. Because of that difference, the identity statements differ in their logical properties and convey different information. In the Wilson problem and the General problem, the difference between sentences lie in their respective hybrid truth-conditions. The hybrid truth-condition of the sentence "Phosphorus is the same size as Hespherus" is that both names refer to the same sized object, which is not the hybrid truth-condition of "Hespherus is the same size as Hespherus". The hybrid truth-condition of "Hesperus is moonless" is that "Hesperus" refers to an object that is moonless, while the hybrid truth-condition of "Phosphorus is moonless" is that "Phosphorus" refers to to an object that is moonless.

In the seventh chapter, Perry responds to Alonzo Church's Slingshot argument for truth-values as references of sentences. The argument states that sentences refer to truth-values because they are the only thing that remains preserved when we either substitute an expression in a sentence with a co-referring expression or when we redistribute parts of sentences, and what remains preserved in substantiation and redistribution is what sentenced refer to. Perry counters the argument using the reflexive-referential theory he developed in the prior chapter. Truth values, contra Church, are not the only thing that remains preserved. In the case of substitution, referential truth-conditions are preserved; in the case of redistribution, hybrid truth-conditions are preserved. So the Slingshot argument gives us no reason to think that truth-values are the reference of sentences.

In the eighth chapter, Perry shows how the ideas from *Beggriffschrift* and the sense and reference theory can be combined into one single framework he calls the Integrative theory. It has three levels of meaning: sense, reference, and extension. The sense is the sense of the sense and reference theory: the sense of names, predicates, and Thoughts. The reference is the reference from Frege's conceptual content theory: circumstances, objects, and properties. The extension is reference from the sense and reference theory: objects, courses of values, and truth-values. Perry enumerates many innovations of the theory. One innovation of this theory is that in indirect discourse and attitude reports embedded sentences behave the same as when they are unembedded, i.e. they refer to the same thing, a circumstance, which instantiates the Thought. So substitution of co-referring expressions in the embedded sentences preserves truth. There is no doctrine of indirect reference, and the Fregean sense is relieved of a burden. Another one is that, because the thesis of unique content is here abandoned, there is a variety of truth-conditions for sentences and expressions that make them up. Further, it gives a better account of predicates, properties, and extensions. What was reference in the sense and reference theory is now extension and, like sense, is unburdened. Perry then gives the truth-conditions of sentences. They are determined by their grammar and meaning. Given this, the reflexive truth-condition of a sentence is that (i) each expression has a sense, (ii) each sense determines a reference, (iii) each reference determines a denotation, and iv) further requirements imposed on these senses, reference, and denotations by the grammatical structure. The referential truth-condition is that there is a circumstance and that the circumstance is a fact. And finally, the truth value of the sentence with its denotations given is determined by the course of values which depending on the truth values of names attaches the same truth value as the extension of the sentence. Then, Perry deals with four potential problems for the Integrative theory that he has to solve since it does not appeal to the doctrine of indirect reference and instead assumes that embedded sentences in indirect and attitude reports refer the same way as when they are unembedded. At this point he still assumes the second thesis that (B) belief is a relation to a proposition. The first three problems he solves in this chapter. The fourth problem he solves in the ninth chapter where he replaces the propositional thesis with the episode thesis. The first problem is intensionality, the problem of explaining the case when substituting co-referring expressions does not preserve truth. The answer is that expressions cannot be substituted though they share the extension because they do not actually co-refer. If Elwood believes that humans are creatures with hearts but does not believe that humans are creatures with a kidney, the embedded sentences about humans do not co-stand for the same circumstance. The second problem is the opacity of descriptions, the problem that the substitution of co-referring description does not preserve truth. The answer is that the descriptions are not co-referring because they refer to different properties despite sharing the same extension. If Elwood knows that the author of Tom Sawyer was born in Missouri but does not know that the author of Huckleberry Finn is born there, then the descriptions do not co-refer. The third problem is the opacity of names and predicates, the problem of explaining the case when substituting co-referring names and predicates does not preserve truth. The answer is the same one for the first and second problem. They might have the same extension but they do not refer to the same thing. If Elwood on his exam marks the claim that Mark Twain wrote Huckleberry Finn as true and marks the claim that Samuel Clemens wrote Huckleberry Finn as false, then this is the reflection of his beliefs. Lastly, Perry delves into the intersection between the Integrated theory and pragmatics. He explains that Integrated theory implicitly assumes that indirect discourse and attitude reports have appeared for two reasons – to pass along information about the agent and to provide an explanation of the agent's actions. For this reason, we as speakers are reluctant to substitute a co-referring expression in such contexts as using this expression could be potentially misleading to a listener. One insight of Frege's sense and reference theory is that the way the objects are presented through expressions that stand for them is of equally important as themselves are. The Integrated theory keeps that insight with the pragmatic explanation of why substitution of co-referring expression is in some situation not allowed.

In the ninth chapter, Perry deals with the fourth problem for the Integrative theory, the problem of logical operations on contents. Perry extends the Integrative theory by adding a mental component to it. Having a belief or other attitude, explains Perry, does not consists of only the relation to a proposition but includes a cognitive state or an "episode" made up of ideas that causes one to make an utterance. To incorporate this into the Integrative theory, he explores the relationship between a cognitive state's content, the ideas that make it up, and the cognitive state's causal role. He presents three insights: One, the content constrains the causal relation between cognitive states and actions. Two, the content has reflexive and referential truth-conditions. And three, the content that motivates action is not referential content, but reflexive content. Following this, he formulates a psychological principle that regulates the causal relation between cognitive states and actions, and so verbal actions, which he calls the fundamental principle of folk psychology and which relies on reflexive content: A desire and a belief will motivate an action of will have a tendency to do so if the belief is made true and the desire is satisfied by the object(s) the notions(s) are of instantiate the property or relation the idea is of, and if the execution of the action will guarantee or at least increase the likelihood that the conditions for satisfaction of the desire will be met if the truth-conditions of the belief are met. Perry then proceeds to apply this episode account on various topics in philosophy of language. He uses it to solve a logical manipulation puzzle, a type of puzzle where an entailment is drawn out from propositions that someone believes, and if he or she is rational, he or she must believe in that entailment. If the austere propositional account is assumed, the rational person must believe all logical consequences of the propositions she believes. But depending on the propositions, this makes the person irrational as the consequences of two or more propositions can be contradictory. This does not happen on the episode account. The entailed content that a rational person believes is limited to the reflexive content of its cognitive states and does not go beyond that. Perry also combines it with David Kaplan's semantic theory of temporal indexicals like "here" and "now" to solve semantic problems with sentences that locate the events they refer in time and change their wording depending on time temporal location of their speakers. Kaplan holds that the meaning of indexicals are determined by characters, functions that bind contexts – agents, times, location, and world. and contents. Perry reinterprets Kaplan's characters as function from utterances and cognitive episodes to contents with various parameters called roles, including agent, time, and location, that are determined by the properties of utterances and episodes. According to this account, what explains the cognitive difference between different sentences that refer to the same event are the episodes that speakers have about them, e.g. if I have the episode that § Now is the time to go to the polling places.§ (Perry's notation for episodes), that with the desire to be a good citizen will move to go to the voting booth today unlike the sentence "November 6, 2018 is election" for which I do not have the corresponding episode. Finally, in the tenth chapter, Perry makes a short recapitulation of the theses he argued for in the book.

I highly recommend John Perry's *Frege's Detour*. The greatest strength of Perry's book is its originality. What Perry did was to take Frege's older, less known theory of meaning that even among Frege scholars was considered to be half-baked and immature, and at best, a stepping stone to his sense and reference theory that made Frege a towering giant in contemporary analytic philosophy of language, and use it to develop a new theory of meaning that is still Fregean in spirit. However, it rejects the basic assumption of Frege, the doctrine of unique content. It is a theory that shows that one can make a workable theory of meaning that does not rely on that postulate. Even if one does not agree with Perry in many points he makes in the book, one must admire the achievement. Another thing that makes the books of great interest is Perry's rereading of Frege's mature articles on sense and reference, which puts a new light on things. For example, he corrects the widely held assumption by Frege commentators that Frege identifies senses with modes of representation (58), and he notes that Frege did not give a valid reason to think that truth-values are reference of sentences (72-73). One caveat is that the book assumes a certain level of knowledge of Frege and general issues in the philosophy of language, so philosophy undergraduates and others with an introductory interest in the philosophy of language will have a harder time following the book. Because of its advanced themes, the readership that will most enjoy this book are philosophers specialized in philosophy of language and Frege scholars. All in all, *Frege's Detour* is a worthwhile book.

MATKO GJURAŠIN
Zagreb, Croatia