Contents

Research Articles

Discussion Note

Book Reviews

Organon F 26 (2) 2019: 186–206 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2019.26201 ISSN 1335-0668 (print)

RESEARCH ARTICLE

Categories and the Language of Metaphysics

Mirco Sambrotta*

Received: 9 May 2018 / Accepted: 2 October 2018

Abstract: The purpose of this paper is to better understand what ontologists are doing when they ask questions about the categories of the world. I will take Cumpa's attempts to find out the fundamental structure of the world as a case-study. In one of his latest paper (Cumpa 2014), he conceives the classical ontological question about the existence of the fundamental categories of the world (what are the fundamental categories of the world?) as a question about the category able to unify the two Sellarsian images of the world: the manifest and scientific images, considered as two different languages. According to him, the only category with such an explanatory power is the category of 'facts' (or 'state of affairs'): the fundamental category of what he calls 'the metaphysical language.' I will argue that if Cumpa takes the latter to be a broader language or framework, in Carnap's terms, common to both the ordinary and the scientific ones, then his proposal turns out to be rather problematic (as they are ultimately 'incommensurable'). On the other hand, if he understands it as external to both of them, then his solution ends up being meaningless and devoid of any cognitive content, with at best a practical character and/or an expressive function.

Keywords: Categorial ontology; sortalism; ontological disputes; scientific image; manifest image.

* University of Granada

✎ Department of Philosophy, Faculty of Philosophy and Humanities, University of Granada, Campus de la Cartuja, 18011 Granada, Spain

✉ mirco.sambrotta@gmail.com

“[T]he tendency represented by the running-up against
the limits of language *points to something.*
St. Augustine already knew this when he said:
What, you wretch, so you want to avoid talking nonsense?
Talk some nonsense, it makes no difference!”
(Wittgenstein L., *On Heidegger on Being and Dread*)

1. Introduction

In a recent paper, Cumpa proposes a new criterion for establishing the fundamental category of the world: ‘the materialist criterion of world-fundamentality’ (Cumpa 2014). According to such a criterion, the fundamental category is that with the greatest explanatory power at the time of reconciling the manifest image and the scientific image of the world. Starting from the well-known Carnapian distinction [see (Carnap 1950)] between questions of existence inside and outside a linguistic framework,[1] I will try to examine Cumpa’s related argument in two different ways.

In the first one, I will interpret Cumpa’s proposal as that of looking for the common fundamental category of both the manifest and the scientific image. In this way, I will consider his categorial question as being asked within a common framework to the two languages (the ‘realistic’ and the ‘scientific’ ones, as he calls them). That is to say, a broader framework in which the category he proposes, the category of ‘facts,’ is a common category shared by both, or at least the only one among the various alternatives proposed able of turning this function. Assuming that Cumpa’s analysis is correct, the category of ‘facts’ will have greater explanatory power, hence a greater epistemic value compared to other categories taken into account. In order to defend such an interpretation of Cumpa’s standpoint, I will try

[1] The notion of ‘framework’ here is quite intuitive: the conjunction of the rules of use of some expressions and the circumstances in which such expressions work. That is, the system of linguistic expressions (key terms like substantives and predicates) and semantic rules (or at least core of rules, constitutive rules) governing those expressions. And, at the same time, the circumstances in which such expressions work.

to point out that it is possible to allow for epistemic values only inside a given framework.

On the other hand, in the second one, I will understand Cumpa's proposal as an effort to find out the fundamental category of the world beyond and outside any framework, trying to answer the ontological question "which category does *really* exist?" in its external reading (according to Carnap's dictates). In this case, the conclusions he reaches play just an expressive function. By this I mean they cannot have any semantic or cognitive content at all (at least a straightforwardly factual content) and at best they can be understood as expressions of commitments to certain language choices. They turn out to be just expressions of commitments to adopt the categorial framework in which a specific category (in this specific case, the category of 'facts') occupies the fundamental level. And this not because of some presumed epistemic values that framework has over others, but rather because of some implicit practical virtues (perhaps, the practical advantages of coping with today's increasingly pressing demand to incorporate scientific expressions with those already in use in ordinary language?). Anyway, I will try to underline how the choice of one framework or another appeal to any epistemic value (as the greatest explanatory power), since epistemic values can be assessed only within a given framework. At most, indeed, one can appeal to some implicit practical virtues, which Cumpa should in this sense make explicit in his inquiry.

The remainder of the paper is divided along these lines: in the next Section, I will summarize Cumpa's solution to the fundamental category problem, drawing attention mainly to his (2014) paper. Then, in Section 3, I will assess his solution from a 'sortalist' point of view. I will present this first analysis of Cumpa's conclusions and lay out my principal worries about that solution (though, perhaps, not decisive). In Section 4, I provide an alternative reading. Following Carnap's well known distinction between 'internal' and 'external' existence questions, I will argue for an 'external' approach to categorial issues (and to ontological claims in general). Although promising respect to the previous one, that alternative does encounter some difficulties and does not avoid to pose some problems to Cumpa's model. Or, at least, it leads to rethinking the issue Cumpa raises in a totally different way. The concluding Section 5 consists of a short recap.

As already mentioned, in what follows, my primary aim will be to provide a concise summary and sympathetic critique of Cumpa's solution. I say 'sympathetic' insofar as I believe he has gotten a great deal in his account of the fundamental structure of the world, making significant and original contributions to this important area of ontology and metaphysics. However, I find the particular solution he develops in (Cumpa 2014) potentially problematic or, at least, not sufficiently developed. While I do not think these concerns are quite as pressing as the ones facing Cumpa's account, they are weighty. Nevertheless, in the end, perhaps the primary lesson for those reflecting on the problem of the fundamental category structure of the world is just that further work may still be needed.

2. The materialist criterion of fundamentality

'Fundamental' is a much debated term in contemporary metaphysics. 'Fundamentality' is also the main concern of Cumpa's work in the last few years.[2] Especially in (Cumpa 2014), he focuses on what he calls "world-fundamentality;" that is to say, the fundamental structure of the world. The question he seeks to answer deals with one of the most classic problems in ontology and metaphysics: "Is our world a world of Aristotle's ordinary substances, Locke's physical substances, Husserl's wholes, Wittgenstein's facts, Sellars's processes, or Quine's sets?" (Cumpa 2014*,* 319). In short, what are the most basic categories that make up our world?

Cumpa suggests that this long-standing dilemma is only possible to be solved by appealing to epistemic values, those in literature are generally labeled as 'theoretical virtues.' Nevertheless, according to him, the traditional epistemic values (or theoretical virtues) usually invoked in metaphysics, such as 'independence' and 'simplicity,' are old-fashioned and fruitless criteria to be used as a guide to find out the most fundamental category of the world. Thereby, he proposes to add a new epistemic value as a criterion of world-fundamentality to the existing catalog of independence and simplicity: the explanatory power. In particular, the explanatory power to

[2] There are of course important issues here as to what we mean by 'fundamental;' on this subject, see (McKenzie 2011, 2014).

account for the relation between 'ordinary world' and 'physical universe.' Therefore, the only categories he thinks can be considered fundamental are those which manage to understand the reconciliation of the ordinary and the scientific description of the world. Or better, he attempts to show that the fundamental categories are just those which have the explanatory power to account for the relation between the ordinary and the scientific image. According to such a criterion, which he calls "the materialist criterion of world-fundamentality," in order to establish whether or not an alleged category can be deemed as fundamental, metaphysicians should consider its explanatory power to account for the relation between the ordinary world and the physical universe.

Next, he argues that the only category which satisfactorily accounts for the relationship between the ordinary and scientific levels of thinghood is the category of 'facts' (or 'state of affairs'). And this leads him to conclude that "the world is a world of facts" (Cumpa 2014*,* 321). In order to demonstrate such an explanatory power of 'facts' to rationally reconstruct the supposed relation, he discusses first some classical alternatives to them as explanatory categories. First, he considers the cases of 'sets' and 'substances,' and he shows why such categories fail to account for the relationship between the two levels, despite the fact that they are usually held to satisfy the traditional criteria of fundamentality (such as 'simplicity,' for instance). Given the division in which the categories at stake are customarily compound, such as substance–accident, set–member or fact–constituent, just the latter has the epistemic primacy to manage to account for the relationship between the ordinary and scientific description of the world. As an example, he takes the 'arrangement of particles' of which a table consists and its 'perceptual properties' as the two constituents of a fact. And, in light of the above outlook, he maintains that just the fact–constituent division can account for the explanatory relationship between the arrangements of elementary particles of the physical universe and the emerging properties of the ordinary world (Cumpa 2014*,* 322).

Since it is not my intent here to question this particular point, I will not go into more detail on this stage of Cumpa's argument, so I will take for granted that the division between facts and constituents has the advantage, over other alternatives under consideration, to possess this cross-sectional character. The issue I am most interested in is the distinction, at the bottom of his view, between the ordinary and scientific levels. What does he exactly mean with 'ordinary world' and 'physical universe?' As he explicitly states, with those expressions he means something similar to what Wilfrid Sellars defined 'the manifest image' and 'the scientific image' [see (Sellars 1963)]. Therefore, by 'ordinary world' he means "an ordinary level of thinghood with which ordinary people are acquainted in their commonsensical and practical experiences" (Cumpa 2014*,* 319). On the other hand, by 'physical universe' he means "a scientific level of thinghood with which scientists are acquainted in their experimental research, such as fundamental physics, chemistry, or biology" (Cumpa 2014*,* 320).

Here, in both Cumpa and Sellars, the background seems to be a unity-of-science view[3] that sees the sciences as forming a reductive explanatory hierarchy, with fundamental physics at the bottom, chemistry built on it, biology on it, the special natural sciences above them, and psychology and the social sciences hovering somehow above them, at least insofar as they deserve to count as 'real' sciences. The ideal is to be able to do all the explanatory work of the upper levels by appeal only to vocabulary and laws of the lower levels.[4]

The alleged fundamental categories of 'facts' should thus account for the world as a complex composed of ordinary objects and the imperceptible objects postulates by fundamental sciences. However, what is more important for the general aim of this paper is that Cumpa clearly considers the source of knowledge of these levels to be respectively the ordinary discourse and scientific theories.

[3] Championed by Neurath and the first Carnap among others, and more recently endorsed by Kim (1992).

[4] Yet today, hardly any philosopher of science would subscribe to the explanatory hierarchy central to the unity-of-science idea. It now seems clear that science works at many explanatory levels, and that generalizations available at one level cannot be replaced by those formulable in the vocabulary of other levels [see especially (Fodor 1974), (Putnam 1975), (Dennett 1991), and (Wilson 2008)]. The explanatory heterogeneity and incommensurability of the various sciences, from which no 'best realizer' emerges, is sometimes called the "Many Levels" view. Thanks to an anonymous referee for pointing this out.

In order to ground the epistemology of 'commonsense realism' and 'scientific materialism,' he accordingly proceeds in the analysis of verbal behavior and scientific laws. What turns out to be at issue are ultimately 'the ordinary language' and 'the scientific language,' or better "the realistic language,"[5] as he calls the former, as opposed to the "the physicalistic language," as he calls the latter (Cumpa 2014*,* 320, 322).

In order to address the question concerning the relations between the descriptions and explanations whose home is in the manifest image and those whose home is in the scientific image (or better, in any scientific images), he conjectures that it is possible to build a cross-sectional language with the explanatory power of reconstructing the two images in one. He trusts in the possibility of 'a metaphysical language' (Cumpa 2014*,* 321) able to display an image of the world as a whole. That is to say, the world composed of the ordinary world and the physical universe. Metaphysical language is not either the realistic language or the scientific language, but at the same time it cannot dispense with both of them. And in this language, the fundamental category is, of course, that of 'facts.' In this way, Cumpa shifts to a special language that smells like the Ontologese and thereby revives hard metaphysical debates.

At this stage, the question I would like to raise is therefore whether the 'metaphysical language' must be taken as a common language to the ordinary and scientific ones, a language which both share (at least at the fundamental categorial level); or instead, it should be better understood as another language different from both of them (to some extent, beyond and outside of both of them). In the next chapter, I will try to develop this concern in the light of the well-known Carnap's distinction between 'internal' and 'external' ontological questions about the existence or reality of entities. Besides, in doing so, I will take a category to be fundamental if and only if it is not derived from another category in a language or framework.

[5] Note that the language of the manifest image (the language of the ordinary lifeworld, both before and after the advent of modern science) does not only deploy normative vocabulary, but also deploys vocabulary to describe and explain.

3. A sortalist reading

The divergence between the world-descriptions provided by physical science and common sense has led to some of the oldest and most persistent arguments for eliminating ordinary objects. For if, as some have thought, the descriptions of science compete with those of common sense, usually the former has primacy over the latter and we must accept that common sense descriptions of the world (as containing trees, battles, and basketballs) apply to nothing. Eliminativism about ordinary objects may seem a radical position to adopt but it is one that meshes with our understanding of contemporary physics, according to which there is only a limited number of certain fundamental kinds of elementary particles and four fundamental forces.

One of the strongest forms that such arguments can take, inspired but apparently not endorsed by the astronomer Sir Arthur Stanley Eddington, alleges not just that the descriptions or claims of physical science compete with those of common sense, but that there is a real conflict between them, a conflict that physical science wins. Thus, if the two are rivals, surely (it is said) the scientific view must win out at the expense of the common sense view, and we must deny the existence of ordinary objects in favor of an ontology sanctioned by physical science. The idea that the descriptions of the world provided by physical science conflict with those of common sense was initially advanced by Eddington's famous discussion of the 'two tables':

> Yes; there are duplicates of every object about me—two tables, two chairs, two pens [...] One of them has been familiar to me from earliest years. It is a commonplace object of that environment which I call the world [...] It has extension; it is comparatively permanent; it is coloured; above all it is substantial [...] Table No.2 is my scientific table [...] My scientific table is mostly emptiness. Sparsely scattered in that emptiness are numerous electric charges rushing about with great speed; but their combined bulk amounts to less than a billionth of the bulk of the table itself. (Eddington 1928, ix–x)

The descriptions of the 'table of science', Eddington emphasizes, do not merely differ from the descriptions of the 'table of common sense', they conflict with it in various ways, e.g. that common sense table is 'substantial' and solid, while the scientific table is "nearly all empty space" (Eddington 1928, x) and so neither substantial nor solid. Quite similarly, Sellars himself maintains that, since each of them purports to be true and complete, any account which attempted to incorporate both the manifest and scientific images "would contain a redundancy" (Sellars 1963, 25). Eddington's attack has been taken up again more recently by Thomasson (2007), who defends an ontology of ordinary objects against eliminativist arguments. According to her, there can be a conflict between them only if the two sides are talking about the same thing. That is to say, in order to demonstrate a conflict one must show that the two descriptions are talking about the same thing with one asserting that it is, say, solid, and the other denying that it is solid. But, Thomasson maintains, any account of what there is presupposes a certain sortal framework. For either side, in order to make a definite claim, must employ some sortal term capable of establishing what is being talked about (and attributed or denied solidity). The sortal which common sense uses (and that Eddington uses) is "table." Nevertheless, it is at least doubtful that scientific theories use sortals such as "table." Susan Stebbing, for instance, famously argued that it is absurd to speak of the object of scientific description as a "table" at all (supposedly in competition with the familiar table) (Stebbing 1937, 54), since scientific objects are mostly 'simples.' We pretty clearly have examples of common sense and scientific discoveries speaking of the same things, in the same terms (and if they are not, the case for a conflict evaporates). However, this is precisely not the case regarding common sense claims about there being tables, apples, and tennis balls, and the claims of contemporary physics couched in terms of waves and particles.

In short, we can define 'sortalism' as the view that highlights the importance of sortal terms and concepts in establishing reference and the truth-conditions of metaphysical claims.[6] In particular, here sortal considerations enter the picture insofar reference to things is fixed via some categorial framework. Hence, Thomasson concludes:

[6] According to Jonathan Lowe (1989), that consists of three claims:

1. Sortal terms and concepts are (generally) associated with semantic principles that supply criteria of application and criteria of individuation and identity for anything that is to fall under them.

> Scientific theories […] do not use sortals such as 'table,' and if science and common sense are using sortals of different categories, the 'things' picked out by the two descriptions cannot be identical. (Thomasson 2007, 142)

Reference is only determinate to the extent that a term is associated with a categorial conception determined by the application and coapplication conditions associated with our terms.[7] In other words, counting claims rely on identity claims, the truth-conditions for which are, she argues, category-relative (Dummett 1973/1981, 74; Geach 1962/1980, 63). Of course, categorial conceptions may be expressed in categorial terms (such as 'organism,' 'artifact,' etc.), which are just highly general sortal terms. And, according to the sortalist view, since the scientific image and manifest image are using sortals of different categories (associated with different application and coapplication/identity conditions), so that they are each concerned with different categories of entities and employ different characteristic sortal terms, we cannot say that the two descriptions conflict with each other. Likewise, we cannot say that there are true identity claims relating the descriptive terms in the vocabulary of the manifest image that refer at all and descriptive terms drawn from the vocabulary (or vocabularies) of the scientific image.[8]

2. Individuals may only be referred to, (re-)identified, and counted by (explicitly or tacitly) employing a sortal.
3. Individuals *a* and *b* can only be identical if they are of sorts with the same criteria of identity, and they meet those criteria.

[7] According to Thomasson (2007, 2009) 'application conditions' are the rules for using nominative terms which establish in what situations they are properly applied, and where they are to be refused; on the other hand, 'coapplication conditions' are the rules for using nominative terms which establish under what conditions we may use the term to refer again to the same entity.

[8] These are in general what we can call 'strongly cross-sortal' identity claims: claims relating terms whose governing sortals are governed by quite different criteria of identity and individuation. But, strongly cross-sortal identities are never true. For the different criteria of identity and individuation associated with the sortals. The claim that strongly cross-sortal identities are never true is a radical one. But, if all that is right, then the relation between the objects referred to in the manifest image and those referred to in the scientific image cannot be identity.

Moreover, since such images are distinguished from each other in terms of the sortal and categorial terms each employs (with the manifest image omitting terms for imperceptible fundamental particles and the like, and the scientific image omitting terms for artifacts, social objects, and the like), they, in fact, do not employ all possible categorial terms. An account can only offer a complete description in terms of that framework in the sense of covering all the things in those categories. The scientific and manifest images presuppose different sortal frameworks and hence they cannot be deemed to be complete in any way that renders those rivals (Thomasson 2007, 148). Consequently, acceptance of the scientific image does not require rejection of the ontology of the manifest one. Therefore, even if each categorial framework purports to be complete in some sense (i.e. offering a complete account of things of those sorts), they still do not purport to be complete in some absolute and 'external' sense.

Of course, conditions of application and/or coapplication for some terms may be built upon others [as, e.g., the conditions for application and coapplication of nation terms may be built upon those for person-terms, landmass terms, etc.; (Thomasson 2009, 451)], making some more basic than others. In this respect, since the manifest image and scientific image employ different characteristic sortal terms, they are each concerned with different categories of entities, and hence with different most fundamental ones. So, even if each categorial framework purports to offer its own account of what the fundamental category of the world is in some sense, they still do not purport to offer its own account of what the fundamental category is in some external and absolute sense.

In sum, the supposed rivalry between scientific and manifest image accounts of what there is can only arise based on the assumption that each image purports to offer (at least in principle) a true and complete account of what there is (Sellars 1963, 20). But, properly understood, neither of the two images (with its own characteristic sortal terms) can really purport to offer a complete account of what there is. Therefore, there is no obvious sense in which either the scientific image or the manifest image may legitimately purport to be complete in a way that would rule out the other. In the same way, each image purports to offer (at least in principle) a true account of what the fundamental category of the world is. However, each image (with its own characteristic sortal terms) can purport to offer a true account of what the fundamental category of the world is in some sense. But, properly understood, neither of the two images (with their own characteristic sortal terms) can really purport to offer a true account of what the fundamental category of the world is in some absolute and 'external' sense.

At this point, one option can be to explore the possibility of meshing the common sense framework with the physics one by constructing some metaphysical relations; another, as we shall see, is to radically remove the necessity for positing certain such relations cleaving them entirely apart, as Thomasson does. According to the first way, the two frameworks are kept in touch with each other. Trying to find a common fundamental category utilized in both scientific and common sense descriptions, Cumpa seems to move exactly in that direction. First, Cumpa dismisses the possibility that, among others, the categories of substance or set are able to achieve this goal. Likewise, Thomasson rejects the possibility to appeal to a common notion of, for instance, 'physical object' or 'occupant of a spatio-temporal region,' insofar the former finds no place within physics itself, and the latter is hardly common in everyday descriptions. Nevertheless, unlike Thomasson who maintains that the conceptual frameworks and ontologies of common sense and physical science are so different that it is hard to find a common conceptual or categorial ground, Cumpa attempts to advance a positive account. Indeed, he argues for the category of 'facts' as able to build such a bridge between the two images (at least according to this first interpretation of his argument). Cumpa's issue then is to establish whether such a relationship effectively holds while neither reducing the common sense framework to the scientific one, nor considering the general metaphysical characterization of such relationships in terms of 'grounding.'[9] To some extent, he takes this relation seriously, metaphysically speaking, without the kind of

[9] Say: *a* is said to be grounded in *b* in the sense that *a* holds in virtue of *b* (without being the case that only *b* exists). Thus, for example, the 'fact' of there being a table in front of me (or Eddington) is grounded in facts about the relevant aggregate of quantum particles in the sense that the former fact holds in virtue of the latter [see (North 2013, 26)].

dependence that 'in virtue of' signifies and he indicates, in at least a preliminary way, how an appropriate metaphysics might be constructed on this basis. Now, explanatory relations, such as the one he outlines, offer a broader framework than, say, causal accounts, whilst not trivializing the relationships as deductive accounts do [see (Thomasson 2007)].

Anyway, endorsing this solution one could face with some problems. As we have seen, claims involving 'facts' (as well as 'physical objects,' 'things,' etc.) are truth-evaluable just if the speaker uses it sortally. And 'facts,' like 'things' or 'objects,' (although it seems to be used non-sortally) is used as a sortal just if it is associated with application and identity conditions outlining what it would take for there to be a fact in a given situation, and under what conditions we would have the same fact again. Clearly, each framework could replace 'facts' with one sortal from its own framework, but then neither is purporting to offer a complete account of 'facts,' but just of 'facts' of that sort. Sortal uses of 'facts' will not help bolster claims to absolute fundamentality either, since, if 'facts' is being used as a sortal (even if it is understood as the fundamental category in that framework) it does not rule out the possibility there being different fundamental categories in other frameworks (for other sortal uses of 'facts'). And besides, if each uses 'facts' in this covering sense that presupposes a different range of sortals, then their resulting accounts of what the most fundamental category is cannot even be true rivals.[10]

In spite of this supposed incommensurability between the two images, Cumpa seems to offer a picture able to retain the category of 'facts' as fundamental and, at the same time, shared by both the realist and the physicalist languages. The dilemma is effectively resolved insofar 'facts' is understood as a compound category which has the highest category of both languages as constituents. In this way, the manifest and the scientific images turn out to be not two different frameworks, but two branches of a broader one which has the category of 'facts' as the most fundamental

[10] It must be noticed that arguments put forward in this Section are also available for any other metaphysical category (e.g. events, processes or states of affairs) insofar as the cross-sectional feature required by the fundamentality mentioned in relation to 'facts' are not met by other metaphysical categories either.

one. That could be a manner of conceiving what he calls 'language of metaphysics.' In this light, Cumpa's proposal could be taken as a viable option and a plausible answer to the original question: "What are the fundamental inhabitants of the world?" Moreover, this approach would also undermine the kind of reductive analysis that physics appears to push us toward. Nevertheless, in order to demonstrate the non-incommensurability of the two frameworks at hand, surely further work needs to be addressed. Complicated issues arise about whether this metaphysical maneuver is really available, but we do not need to address them here, for even if such a move is possible, it will help revive neither a rivalry nor compatibility between them, strictly speaking.[11]

4. The external reading

As we have seen, the sortalist position gives us reason to doubt that each of the two images could legitimately purport to provide an account of what the fundamental category *absolutely* is. Since each image (with its own characteristic sortal terms) purports to offer its own account of what the most fundamental category is in some sense, we cannot legitimately say they provide rival accounts of what the fundamental category is. Nevertheless, there is at least another possible interpretation of Cumpa's project. Employing Carnap's terminology, I will call it 'external reading.' Indeed, one might try to present the conflict in terms of some neutral sense of 'facts,' external to any framework that establishes the rules of use for such a term. But 'facts,' in that sense, would not then be a sortal term and could

[11] A related worry is that, even if a category that covers all possible (first-order) categorial concepts is possible, set-theoretic-style paradoxes, such as a Russell-style paradox, quickly arise. We can postulate a category that covers all possible (first-order) categorial concepts ('organism,' 'artifact,' etc.) and all of their compliants, but then there are possible (second-order) categorial concepts which are not covered (e.g. first-order category), so there is a sense in which we have not covered absolutely universally. So that it seems there is no category of which one could rightly claim to be absolutely universal. But more than that, it seems that we can "form no definite conception of the totality of all objects which could be spoken of" (Dummett 1973/1981, 566–67, 582–83).

not be used to establish reference. That is, if ‘facts,’ in its neutral use, is not a sortal term, then, on the sortalist view, it cannot enable us to establish reference to something, about which science and common sense may then agree or disagree. Consequently, we cannot legitimately say that ‘facts’ is the fundamental category of the world, where ‘facts’ is being used neutrally. For if ‘facts’ is not being used as a sortal term, it does not come associated with application conditions needed to establish if it is properly applied and the identity criteria (coapplication conditions) needed for counting. Thence, we have serious reason to doubt that such alleged neutral uses of ‘facts’ could be used to answer the question about what the fundamental category of the world is. The question "is ‘facts’ the fundamental category of the world?", understood externally (external to any framework), turns out to be an ill-formed, unanswerable question. Likewise, claiming that "‘facts’ is the fundamental category," so understood, will also result meaningless and devoid of any cognitive content. In sum, if ‘facts’ is really used neutrally in attempts to state these debates, then that should raise our suspicions that the claims involved are incomplete and not truth-evaluable. In the same way, that should raise our suspicions that the corresponding metaphysical questions are ill-formed and unanswerable, and that apparently competing answers to them do not really conflict with each other.

Nevertheless, even though they so understood result to be cognitively meaningless and fail at bipolarity (they have no true values), they may still have a different sort of ‘meaning:’ a normative one. Indeed, the statement "the fundamental category is that of ‘facts,’" in its external use, may express the commitment to adopt a framework in which ‘facts’ occurs as the fundamental categorial term (in that particular framework).[12] And, perhaps, such a framework could be identified with what Cumpa calls ‘the metaphysical language.’ Anyway, this external use says nothing about that framework itself, what actually it is, how it is constituted and whether it is a possible language at all. Moreover, if the ‘metaphysical language’ is taken to be different from both the realistic and the scientific language, it will be deprived of any relationship with them, and to a certain extent, it will be

[12] For an expressivist account of ontological claims and questions, taken externally, see (Kraut 2016).

incommensurable with both of them. Thus, 'facts,' understood as the fundamental category of the metaphysical language, will certainly not play that role also in the other languages at stake. However, if Cumpa has in mind some kind of relationship (even some kind of metaphysical relationship) between the alleged metaphysical language and the two other mentioned, I think he should make it explicit, specifying or at least clarifying the supposed contact point.

Furthermore, if this is effectively the most reliable interpretation of Cumpa's proposal, appealing to epistemic virtues (as Cumpa suggests when he argues for the greatest explanatory power of 'facts') does not seem to be a possible strategy to be followed. According to the present view, no framework can be deemed more correct or valid than any other. Or better, since speaking of correctness (or validity) here does not apply at all, then it is applied in the same way. Likewise, among the frameworks, there is none that is uniquely best (viz. the 'correct' one). But this formulation certainly does not suggest that the frameworks are all equally good: definitely, a framework might be better than another according to some goals. The linguistic rules we adopt need not be arbitrary, given our purposes, since some rules may serve those purposes better than others. Some languages may be better than others for various purposes and there may be practical issues, or reasons, involved in determining which language is better for that given purpose (or set of purposes). Hence, it follows that virtues for opting for one language over another cannot be epistemic but at most practical in character.

It is also important to notice that, insofar as such practical virtues (or non-epistemic values) act like norms or standards of evaluations, these comparative judgements, of which frameworks are better than which, turn out to be normative. Or in other words, even when based in part on non-normative descriptions, they can only be made from those norms. Therefore, such judgements of betterness must be understood as involving a hidden relativity to a norm; in particular, some practical value or virtue. In this sense, it may be quite reasonable to engage in debates about the merits of these various proposals, practical proposals about which set of concepts (or revisions of our current concepts) would best serve some particular set of purposes, though it would be misguided to think of these as substantive debates about how the world is actually made up.

This reading is very close to how Carnap suggests we should understand external ontological questions in general: as practical questions about the advisability of adopting certain linguistic forms. Although, according to Carnap, external questions have no cognitive content at all, they are still significant questions. Indeed, they are not meant to be questions about what there is in the world, but rather questions about what we should do: questions about which framework we ought to use according to some practical goals. Correspondingly, ontological claims, taken externally, are to be conceived as implicitly answering practical questions about whether or not to accept the related linguistic framework as a whole. And those, of course, are quite different from the (internal) cases in which "we have to make the choice whether or not to accept and use the forms of expression in the framework in question" (Carnap 1950, 207). Therefore, the relevant distinction turns out to be the one between the theoretical issues about what true statements (including existence claims) may be made using a given linguistic framework and the purely practical issues of which linguistic frameworks to choose and adopt. And the choice of a language is nothing but a purely practical choice about what tool to use, rather than as a theoretical decision that is either correct or incorrect: "it does not need any theoretical justification because it does not imply any assertion of reality" (Carnap 1950, 214). In short, if we take external categorial questions literally (as attempted theoretical or factual questions), they are ill-formed pseudo-question. The best we can do is then to consider them as implicitly asking questions about whether or not to accept and use a given categorial framework (with its own categorial structure and fundamental categories).

But, Cumpa does make no reference at all to the practical purposes for which such a metaphysical language should be adopted. Might these be, perhaps, the practical advantages of coping with today's increasingly pressing demand to incorporate scientific expressions with those already in use in ordinary language? Anyway, if that is precisely how Cumpa intends the role of the claim that "'facts' is the fundamental category of the world" and the function of 'metaphysical language' in general, then, I guess, he should at least mention them, as long as it is possible. In that direction, in order to reveal what they actually might be, further investigations are certainly still needed.

5. Conclusion

The distinction between structure and content is one that has arisen repeatedly in discussions over the relationship between the scientific and the everyday ontology, but it evaporates as far as Eliminativism is concerned, since all relevant content is taken to be cashed out in structural terms. However, according to Thomasson, Eddington's standpoint is undermined because, she claims, there is a "lack of conflict between the merely structural properties physics imputes to the world and the qualitative content involved in ordinary world descriptions" (Thomasson 2007, 139). Insofar as the two manifest and the scientific images involve different linguistic/categorial frameworks, we are not in a position to compare them and then it would be a mistake both to maintain that there is and that there is not a conflict between them.[13]

Cumpa (2014) adopts a different strategy. He argues neither for the incommensurability of the two languages nor for the reducibility of the ordinary level to the scientific level of thinghood. Instead, he attempts to find a category able to reconcile the two images. He identifies the category of 'facts' as the only one which meets this requirement: the best category to account for the relation between the ordinary world and the physical universe. As he defines it: "The fundamental category of the world." Nevertheless, it turns out to be not clear at all how he suggests the relationship between the alleged category of 'facts' and the two descriptions of the world ought to be understood.

The aim of this paper has been to outline two possible ways in which Cumpa's factualist approach could be conceived. Both, however, present some difficulties, or at least they need further investigations. According to

[13] In the same spirit, the general view I have been elaborating and defending in this paper is that many manifest-image descriptive expressions which scientific naturalists have relegated to second-class citizenship in discourse are not inferior, just different. It just is not the case that everything we talk about in the manifest image that exists at all is something specifiable in the language of an eventual natural science and that "in the dimension of describing and explaining the world, science is the measure of all things, of what is that it is, and of what is not that it is not" (Sellars 1956, §41).

the first one, common sense image and scientific image are taken to be two branches of a single broader linguistic framework, which he calls ‘metaphysical language.’ Along these lines, ‘facts’ turns out to be the fundamental category in that language and, as such, a category shared by both images. Nevertheless, rather than a category common to the realist language and the scientific language, ‘facts’ is considered to be a compound category, which has the highest category of both (‘arrangement of particles’ and ‘perceptual properties’) as constituents. In other words, ‘facts’ should be understood as the fundamental category of a broader framework (the metaphysical language), but at the same time as constituted by the highest categories of both those narrower frameworks (the realist language and the scientific language). Appealing to a sortalist standpoint, in Section 3, I have tried to reveal the limits of this way of conceiving Cumpa’s proposal.

Alternatively, in Section 4, I have introduced what I called an ‘external’ reading. Here, evoking Carnap’s well known distinction between ‘internal’ and ‘external’ ontological questions, I have presented Cumpa’s claim that “‘facts’ is the fundamental category of the world” as external to any linguistic/categorial framework and the term ‘facts’ as used in some neutral sense (as a non-sortal term). I have argued that such an external categorial statement is meaningless as devoid of any cognitive content. Following Carnap, I have suggested that at best it might be understood as a normative claim. That is, not as a descriptive claim, but rather as a claim about what we should do. In particular, a statement about what categorial framework we ought to adopt. In this respect, it will express commitments to the adoption of a categorial framework in which the fundamental category is that of ‘facts.’ Then, I have tried to show how such a reading clashes in principle with Cumpa’s conception of a ‘metaphysical language.’

In conclusion, whichever of the two interpretations is closer to Cumpa’s original purpose, further explanations and clarifications, I think, are needed. I hope Cumpa is willing to take up my suggestions and to address these issues developing his account in one direction or another.

Acknowledgements

I would like to thank Javier Cumpa for his helpful comments on this paper.

References

Carnap, Rudolf. 1950. “Empiricism, Semantics, and Ontology.” *Revue Internationale de Philosophie* 4: 20–40. Reprinted as a supplement to: Carnap, Rudolf. 1956. *Meaning and Necessity: A Study in Semantics and Modal Logic*, enlarged edition. Chicago: Chicago University Press.

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

Dennett, Daniel C. 1991. “Real Patterns.” *Journal of Philosophy* 88 (1): 27–51. https://doi.org/10.2307/2027085

Dummett, Michael. 1973/1981. *Frege: Philosophy of Language*, 2nd edition. Cambridge, MA: Harvard University Press.

Eddington, Arthur S. 1928. *The Nature of the Physical World*. Cambridge: Cambridge University Press.

Fodor, Jerry. 1974. “Special Sciences: The Disunity of Science as a Working Hypothesis.” Synthese 28 (2): 97–115. https://doi.org/10.1007/BF00485230

French, Steven. 2014. *The Structure of the World: Metaphysics and Representation*. Oxford: Clarendon Press. https://doi.org/10.1093/acprof:oso/9780199684847.001.0001

Geach, Peter. 1962/1980. *Reference and Generality*, 3rd edition. Ithaca, NY: Cornell University Press.

Kim, Jaegwon. 1992. “Multiple Realizability and the Metaphysics of Reduction.” *Philosophy and Phenomenological Research* 52 (1): 1–26. https://doi.org/10.2307/2107741. Reprinted in: Kim, Jaegwon. 1993. *Supervenience and Mind*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511625220.017

Kraut, Robert. 2016. “Three Carnaps on Ontology.” In *Ontology after Carnap*, edited by Stephan Blatti and Sanda Lapointe, 31–58. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199661985.003.0003

Lowe, E. Jonathan. 1989. *Kinds of Being: A Study of Individuation, Identity and the Logic of Sortal Terms*. Oxford: Blackwell.

McKenzie, Kerry. 2011. “Arguing Against Fundamentality.” *Studies in History and Philosophy of Science, Part B* 42 (4): 244–55. https://doi.org/10.1016/j.shpsb.2011.09.002

McKenzie, Kerry. 2014. “Priority and Particle Physics: Ontic Structural Realism as a Fundamentality Thesis.” *British Journal for the Philosophy of Science* 65 (2): 353–80. https://doi.org/10.1093/bjps/axt017

North, Jill. 2013. “The Structure of a Quantum World.” In *The Wave Function*, edited by David Z. Albert and Alyssa Ney, 184–202. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199790807.003.0009

Putnam, Hilary. 1975. “Philosophy and Our Mental Life.” In *Mind, Language and Reality*: *Philosophical Papers*, vol. II, edited by Hilary Putnam, 291–303. Cambridge: Cambridge University Press.
https://doi.org/10.1017/CBO9780511625251.016

Sellars, Wilfrid S. 1956. “Empiricism and the Philosophy of Mind.” In *Minnesota Studies in the Philosophy of Science*, vol. I, edited by Herbert Feigl and Michael Scriven, 253–329. Minneapolis, MN: University of Minnesota Press.

Sellars, Wilfrid S. 1963. “Philosophy and the Scientific Image of Man.” In *Science, Perception, and Reality*, edited by Robert Colodny, 35–78. London: Routledge & Kegan Paul.

Stebbing, Susan. 1937. *Philosophy and the Physicists*. London: Methuen and Co.

Thomasson, Amie. 2007. *Ordinary Objects*. New York: Oxford University Press.
https://doi.org/10.1093/acprof:oso/9780195319910.001.0001

Thomasson, Amie. 2009. “Answerable and Unanswerable Questions.” In *MetaMetaphysics*, edited by David Chalmers, Ryan Wasserman, and David Manley, 444–71. Oxford: Oxford University Press.

Wilson, Mark. 2008. *Wandering Significance*. Oxford: Oxford University Press.
https://doi.org/10.1093/acprof:oso/9780199269259.001.0001

Organon F 26 (2) 2019: 207–227
https://doi.org/10.31577/orgf.2019.26202

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

RESEARCH ARTICLE

Animalism and the Vagueness of Composition

Radim Bělohrad*

Received: 13 June 2018 / Accepted: 2 October 2018

Abstract: Lockean theories of personal identity maintain that we persist by virtue of psychological continuity, and most Lockeans say that we are material things coinciding with animals. Some animalists argue that if persons and animals coincide, they must have the same intrinsic properties, including thinking, and, as a result, there are 'too many thinkers' associated with each human being. Further, Lockeans have trouble explaining how animals and persons can be numerically different and have different persistence conditions. For these reasons, the idea of a person being numerically distinct but coincident with an animal is rejected and animalists conclude that we simply are animals. However, animalists face a similar problem when confronted with the vagueness of composition. Animals are entities with vague boundaries. According to the linguistic account of vagueness, the vagueness of a term consists in there being a number of candidates for the denotatum of the vague term. It seems to imply that where we see an animal, there are, in fact, a lot of distinct but overlapping entities with basically the same intrinsic properties, including thinking. As a result, the animalist must also posit 'too many thinkers' where we thought there was only one. This seems to imply that the animalist cannot accept the linguistic account of vagueness. In this paper the author argues that the animalist can accept the linguistic account of vagueness and retain her argument against Lockeanism.

* Masaryk University
✎ Department of Philosophy, Faculty of Arts, Masaryk University, Arna Nováka 1, 602 00 Brno, Czech Republic
✉ belohrad@phil.muni.cz

Keywords: Animalism; Lockeanism; partial identity; personal identity; supervaluations; vagueness.

1. Introduction

Animalism is a theory of personal identity according to which you and I are animals. A further claim that both some animalists and their opponents make is that animals persist by virtue of biological continuity.[1] For an animal to exist, there has to be a life, and for an animal to continue existing, there has to be a continuing life (Olson 2007, 29; van Inwagen 1990, 145; DeGrazia 2005, 51–56).

There are other popular accounts of how we persist. Historically, the most influential have been psychological theories inspired by Locke, according to which we are not animals, but persons—essentially intelligent beings with sophisticated mental capacities, which persist by virtue of psychological continuity.

At first glance, there does not seem to be a disagreement between these two accounts.[2] After all, could intelligent beings with sophisticated mental capacities not be animals? And could animals not be intelligent and have sophisticated mental capacities? In other words, could a particular animal

[1] Not all animalists accept this claim. See, for instance, Snowdon (2014) for a theory according to which the criterion of personal persistence is the retention of the life-apt structure, or McDowell (1997) who defends the claim that psychological continuity is the criterion of persistence of animals. Also, as Olson (2015) points out, this claim is quite independent of the first claim. Olson coins the conjunction of the two claims *strong animalism* and concedes that it is *strong animalism* that usually stirs up debate. In what follows, when I refer to animalism, I always mean the conjunction of the two claims.

[2] But, as Olson (2015) points out, the two accounts respond to very different questions. Animalism is a response to the question of what we are. Lockeanism responds to the question of how we persist. Neither answer has any direct implications for the other question. Olson also shows that the implications of Lockeanism for the 'What are we?' question are quite unclear. In this paper I accept the minimum that is usually accepted: persons are *material entities* with *complex psychological properties*, such as self-reflection, and, importantly, they are not animals.

and a particular person not be one and the same thing? According to the Lockeans, they could not, because animals and persons differ in their essential properties and in what they can survive. While the persistence conditions of animals are biological, ours are psychological. This means that we persist as long as there is psychological continuity—an uninterrupted chain of mental states.[3] If the chain is interrupted, we cease to exist, even though a living animal may continue to exist (Shoemaker 1984; Lewis 1976).

The claim that the lives of persons and animals may come apart is usually supported by thought experiments involving brain state transfer devices, teleporters, brain transplants, etc., or real-life examples of humans in a coma or a permanent vegetative state [see (Parfit 1984, 197–200), for instance].

There are a number of different psychological theories, but the most widely held ones claim that persons are material entities which are not animals, but are related to animals by a very intimate relation. In some versions this relation is material coincidence—the sharing of matter (Shoemaker 1984, 113). In others it is the relation of constitution (Baker 2000; Johnston 1987, 2007), though this also entails material coincidence (Baker 2000, 43). Just as a statue is constituted by a lump of clay but is not identical to it, because the lump can persist through changes that the statue cannot, a person is (it is claimed) constituted by an animal but, for similar reasons, not identical to it.

Psychological theories of this sort must explain two mysteries.

Mystery 1

If persons are material entities, if each person is made of the same matter as an animal and if persons are not identical to animals, then where I am right now, there are two material entities—a person and an animal—that share every particle of matter and their overall structure. As a result, it

[3] For instance, Parfit defines psychological continuity as the overlapping chains of strong psychological connectedness. Psychological connectedness is the holding of direct psychological connections, such as the connections between memories and the experiences that caused the memories, intentions and experiences of actions resulting from the intentions, etc. Connectedness is strong if ‘the number of connections, over any day, is *at least half* the number of direct connections that hold, over every day, in the lives of nearly every actual person’ (Parfit 1984, 205).

seems impossible to explain plausibly how these two entities could differ in the way that the Lockeans claim.

First, each such animal must be a person. Each such animal has a brain just like the person, the brain is functioning and, if consciousness and thought are generated by the functioning of the brain, each such animal is conscious and thinks. In fact, each such animal has mental capacities as complex as the coincident person. That seems to qualify it as a person, too (Olson 1997, 100). So if I am not an animal, as the Lockeans claim (that is, if I am a person and persons are not animals), there must be two conscious and thinking beings where we normally thought there was only one—me and the animal—which makes the Lockean theory inconsistent, because then there are two persons, only one of whom persists by virtue of psychological continuity (Olson 1997, 106–109).[4]

Second, each such person must be an animal. Being a material entity coincident with an animal, it must be composed of particles that together give rise to a life; it has a heart, lungs, metabolism, immune system, etc. In fact, it has all the characteristics that make the animal an animal. That seems to qualify it as an animal, too.

If an entity's intrinsic properties depend on its microphysical structure, it seems obvious that animals and persons must have the same intrinsic properties. That makes it a complete mystery how one can have biological persistence conditions while the other has psychological ones. How could something make one cease to exist while not affecting the other (Olson 1997, 98)? Call this the metaphysical mystery.[5]

Mystery 2

Suppose we can explain Mystery 1, and where I am there are two material entities that are indistinguishable in terms of their intrinsic properties. Which one of them am I? The question would not be worth answering if the alternative answers had no practical implications. But if I am an animal, I will still exist when in a permanent vegetative state (PVS), whereas if

[4] This problem is often referred to as 'the thinking animal problem' or 'the too many thinkers problem'.

[5] This is a special instance of the general problem of metaphysical grounding. See, for instance, (deRosset 2011).

I am a person, I will not—no person can be identical to any being in PVS, because beings in PVS are not persons and all persons are persons essentially. Further, if I am a person, it may be rational for me to have my head removed at the moment of my death and preserved by Alcor in the hope of it being transplanted onto a new body one day—psychological continuity with the person in the new body will ensure it is me. If I am an animal, this would be an unjustifiable waste of resources, because the animal I am does not get transferred to the new body.[6] But how am I to find out? I may think that I am the person, but the animal thinks the same. One of us is mistaken, though, and there is no way we can find out (Olson 1997, 106). Call this the epistemic mystery.

These two mysteries lead to the following argument against Lockeanism and in favour of animalism: if a particular animal and a particular person completely overlap and share every particle of matter, then they must be indistinguishable in their intrinsic properties. But if they are indistinguishable with respect to every intrinsic property and if we plausibly suppose that persistence conditions are grounded in intrinsic properties, then they cannot differ in their persistence conditions. But then what grounds do we have for claiming that the person is not the animal? If the person shares every intrinsic property with the animal and begins and ceases to exist at the same time as the animal, why suppose it is numerically distinct from the animal? The animal must be as conscious as the person is, the person must have a heart and breathe as much as the animal, and surely it would be absurd to suppose that all of these properties are duplicated where the animal and the person coincide. And even if they were duplicated, it would still be hard to explain the alleged difference in persistence conditions. Thus, the Lockean claim that persons are not animals seems hard to justify. The animalist concludes that where I am, there are not two entities that are both living, conscious, thinking and meet the criteria for personhood. There is just one—the animal. This conclusion seems to be in accordance with both common sense and the findings of the natural sciences.

[6] This claim is not accepted by all animalists. Van Inwagen (1990, 170) argues that head transplants preserve biological continuity and, thus, move the animal to the new body.

2. The vagueness of composition

However, the reasoning that the animalist uses to undermine psychological theories seems to lose its footing when we attempt to make sense of the phenomenon of vagueness of composition.

When examined closely, the human organism, or any organism for that matter, resembles a cloud. (Animalists sometimes actually liken a life, the concept they use to define organisms, to a storm.) In a cloud, the sharp boundaries that we observe from a distance become blurry once we inspect it closely. The further we go from the core of the cloud, the more frequently we will find water droplets that are less and less integrated into the body of the cloud, until we find droplets of which it is impossible to say whether they are parts of the cloud or not.

Organisms are very similar. There are particles that are deeply incorporated into their metabolic system, particles that are clearly not and particles that are in various stages of incorporation, making it impossible to determine whether they are parts of the organism or not. These may include, for instance, particles in molecules of food that are being absorbed into the blood stream or particles in dying skin cells. As a result of this indeterminacy of parthood, it is impossible to determine precisely where the boundaries of an organism lie.

The vagueness of composition affects virtually all material objects we encounter every day. But why is it supposed to be a problem, and why should animalists be especially concerned about it?[7] There is a famous argument, formulated independently by Peter Unger (1980) and Peter Geach (1980), which shows that if we accept the existence of vague boundaries of objects, we are driven to the conclusion that where there seems to be a single object, there are actually a great number of them—something that flies

[7] It is not just animalists who should be concerned about vagueness. Adherents to the bodily view, the brain view and even those Lockeans who believe that persons are material entities (see below) should have an account of vagueness. But it has been animalists who have built their opposition to Lockeanism around the idea that positing two numerically distinct but completely overlapping entities leads to a number of problems. Anyone who says that should be especially concerned about vagueness, as I show below.

in the face of common sense and seems to undermine the animalist's belief that where I am, there is just one material entity that is conscious, intelligent and thinking.

The argument was termed the *Problem of the Many* by Unger. I will now present a version tailored to my purposes.

1. Suppose there is a human organism with vague boundaries. Call it O.
2. Then there are particles D_1–D_n, which are determinately parts of O, and particles I_1–I_n, which are indeterminately parts of O.
3. Then there are various sets of particles S_1–S_n, which are equally suitably arranged to compose[8] an organism. For instance, $S_1 = \{D_1$–$D_n\}$, $S_2 = \{D_1$–$D_n, I_1\}$, $S_3 = \{D_1$–$D_n, I_2\}$, etc.
4. For each set S, the members of S compose an entity. (It would be unjustified to claim that the members of S_1, for instance, compose something, while the members of S_2, differing only by a single particle, do not compose anything. The difference of a single particle seems to be compositionally negligible.)
5. If O exists, then each of these entities is an organism. (It would be unjustified to claim that the members of S_1 compose an entity which is an organism, while the members of S_2, differing only by a single particle, compose an entity which is not an organism. An organism cannot differ from a non-organism by a single particle only.)
6. Then for each organism O, there are a great number of organisms which almost completely overlap O. (Some differ by a single particle, others by more particles, but still negligibly.)

The reason animalists should be concerned about this argument should now be obvious. In their challenge to Lockeanism, animalists show the troublesome implications of Lockeans postulating materially coincident but numerically distinct entities—if a person and an animal coincide, they must share all of their intrinsic properties, they must both think, be intelligent, and

[8] Simply put, composition is a relation among objects such that if the objects O_1–O_n stand in that relation, there is an object P that has objects O_1–O_n as parts.

be persons. As a result, there are more persons and thinkers than we thought there were. Moreover, only some persist by virtue of psychological continuity. However, if animals are vague objects, the animalist will have to face problems of a very similar type. All the different sets contain particles that have, as far as we can tell, an equal claim to compose an animal, and since they differ so minutely, they will all presumably have virtually the same intrinsic properties. Specifically, if I am conscious, they will almost certainly be conscious, if I am thinking and intelligent, they will almost certainly be thinking and intelligent, and if I am a person, they will almost certainly be persons.

It seems that the animalist is driven to a difficulty quite similar to the one he ascribes to the Lockean. If he has to accept a multiplicity of animals in virtually the same place, why could there not be two material entities—an animal and a person—in the very same place? Surely, a few particles cannot make that much difference.

3. Solutions to the Problem of the Many

Metaphysical theories offer a number of solutions to the Problem of the Many. Peter Unger originally suggested a solution that has become known as *nihilism*. It is based on the idea that it is absurd to conclude that all of the sets S_1–S_n contain particles that compose something, and thus denies premise 4. But the only alternative is that none of the sets compose anything. And if none of them compose anything, there aren't any vague objects. That would not be so troubling if vagueness did not affect all of the objects that we encounter in the world. Since it does, where we thought there were vague objects, there turn out to be only particles that compose nothing. As a result, none of the ordinary things we think there are actually exist (Unger 1980, 462).

This solution is radical and certainly not in tune with what the animalist wants to say. Animalism holds that there is at least one sort of composite object—animals. (Whether there are chairs, clouds or rusty nails is a matter the animalist *qua* animalist does not attempt to answer.) Nihilism offers too few entities to the animalist.

There are, however, solutions to the problem that the animalist can accept. One solution favoured by two prominent animalists appeals to ontic vagueness (van Inwagen 1990, 213–227; Olson 2008, 42). This solution claims that the many organisms are a single organism with vague boundaries. Arguably, the animalist can also accept the brutalist response as defended by Markosian (1998), according to which there is no interesting answer to the question of when composition occurs and compositional facts are *brute facts*. On this account, particles in only one of the sets compose an organism, but there is no interesting explanation as to why the particles in the other sets do not compose anything. But I want to look more closely at solutions that do not seem to be available to the animalist because they posit far more entities than the animalist is willing to accept. These solutions are, respectively, the semantic account of vagueness (semanticism) based on the notion of supervaluations, and a solution by means of partial identity as defended by David Lewis. At first sight, the animalist cannot treat the vagueness of animals as a matter of semantic indecision, because that solution requires there to be many equally suitable candidates in nearly the same place at the same time. Lewis' analysis of vagueness of composition also entails the existence of a multiplicity of candidates. If correct, these views would seem to be inconsistent with there being just one animal in the space where I am located.[9]

However, both of these solutions are based on the idea that although the multiplicity is real, we can (in a manner of speaking) "cheat" and pretend that there is just one entity in the place where we want it. The key question is whether the animalist can cheat too.

4. Cheating I

The semantic solution to the Problem of the Many is based on rejecting premise 5 of the argument. According to that premise, the objects composed by the many sets of particles are all organisms. Semanticism denies this—the sets of particles do compose entities, but these entities are mere candidates for being an organism, and only one candidate is an organism [see e.g.

[9] See, for instance, (Zimmerman 2008, 30).

(Lewis 1999, 171);[10] see also (McGee and McLaughlin 2000); (McKinnon 2002)]. Let us look at the details.

Semanticism is a theory according to which vagueness is essentially a linguistic matter resulting from the fact that our expressions do not have precisely specified denotata. According to semanticism, the world is perfectly discrete, containing only entities with sharp boundaries, but our expressions are sometimes indeterminate regarding which of the sharp objects they actually denote. Thus, the expression “tall person” is vague, because it could denote the set of people who are at least 180 cm tall or the set of people who are at least 190 cm tall, and we have never needed to determine the denotatum precisely. The vagueness of a term is thus explained by postulating a number of candidate denotata that are precise (so-called “precisifications”) and saying that it has not been determined which of the precise candidates is the actual denotatum of the vague term.

But sometimes we could, if we wished to, make a vague term precise by selecting out of the many alternative precisifications the one that will from now on be the denotatum. Sometimes we do just that. For instance, we need to precisify the term “the moment of death” for legal and medical purposes. On other occasions it would be entirely pointless to do so, because what we want to say using the vague term will be true regardless of which of the precisifications is the actual denotatum of the vague term.

But the general assumption of the semantic account is that there are many precisifications for a vague term, each of which is a suitable candidate to be the denotatum of the term. The multiplicity of the candidates seems to be inconsistent with the above argument for animalism and against Lockeanism. Suppose there are a number of candidates to be the organism that I am. Although the candidates do not completely coincide, they overlap so extensively that a great number of them will very likely share many intrinsic properties. Specifically, if I am conscious, many others will be. If I am thinking, many others will be thinking as well. As a result, there would seem to be too many thinkers where we thought there was only

[10] Lewis endorses both views to be discussed here—semanticism as well as partial identity. See (Lewis 1999, 179–82).

one—something the animalist certainly does not want to admit in the light of her argument against Lockeanism.

However, it can be argued that the animalist can accept the semanticist framework if it is enriched by a method known as supervaluation, which assigns truth values to statements containing vague terms [see (Lewis 1999, 171–75)]. In this method, each sentence containing a vague term is broken down into many different interpretations in which the vague term is replaced by one of the precisifications, and the statement is then evaluated. Statements which come out true on every precisification are *supertrue*. Statements which come out false on every precisification are *superfalse*. And statements which come out true on some precisifications and false on others are *supertruth-valueless*. The important lesson is that according to this version of semanticism, the goal of communication is to convey not truth, but supertruth.

Accepting the supervaluationist framework gives us the resources to solve the Problem of the Many in a manner consistent with animalism. More specifically, we can show that it is entirely legitimate to say that where I am, there is just one organism.

To see that, let us focus on a particular organism; call it R. Suppose, then, that there are various precisifications of R. Let us assume for simplicity that there are just four:

P1: R is composed of the set of particles $S_1 = \{o, p, q, r\}$;

P2: R is composed of the set of particles $S_2 = \{o, p, q, s\}$;

P3: R is composed of the set of particles $S_3 = \{o, p, r, s\}$;

P4: R is composed of the set of particles $S_4 = \{o, q, r, s\}$.

Suppose that R is now sitting in the living room and watching TV, and consider the sentence "R is watching TV".

The supervaluation of this sentence will assign truth values to all of the precisifications of the sentence which result from replacing R with a precisely defined term. For brevity, let us assume that "R(P1)" means "R under the precisification P1".

The following will now hold:

"R(P1) is watching TV" is true.

“R(P2) is watching TV” is true.

“R(P3) is watching TV” is true.

“R(P4) is watching TV” is true.

Since the sentence “R is watching TV” is true on all precisifications, it is supertrue, and we are entitled to assert it.

By contrast, let us now consider the sentence “R is composed of {o, p, q, r}”. This sentence will be true if R = R(P1), but it will be false on all other precisifications. This discrepancy in truth values results in the sentence being supertruth-valueless. The same goes for “R is composed of {o, p, q, s}”, etc. Since R is a vague term, we cannot claim that its denotatum is determinately composed of a particular set of particles. Any such claim will lack supertruth value.

But now consider the sentence “Only one set of particles composes R”. What will be the supertruth value of this sentence? The above examples suggest that on each precisification only one set of particles composes R. On R(P1) it is S_1 and no other set, on R(P2) it is S_2 and no other set, etc. Since it is true on every precisification, the sentence “Only one set of particles composes R” is supertrue and we are justified in asserting it.

The above considerations entail the following claims: R is a vague term. There are many alternative precisifications of R. It is impossible to say which precisification is the sole legitimate denotatum of R, for none are. However, on any precisification there will be just one R. That means that where I am, there is just one organism, even though it cannot be determinately stated which of the alternative precisifications it is.

Since, according to supervaluationism, the goal of communication is to convey information that is supertrue, and since the sentence “Where I am, there is just one organism” will be supertrue on the supervaluationist account, the animalist is perfectly justified in asserting it.

5. Cheating II

The semantic solution to the Problem of the Many is based on rejecting the claim that all of the entities composed by the different sets are organisms.

Another solution to the Problem of the Many is offered by David Lewis (Lewis 1999, 177–79). Lewis introduces the concept of partial identity to show that although the entities are all organisms, we can still (in a manner of speaking) say that there is just one organism. Because this solution accepts the multiplicity of organisms, it too should be unavailable to the animalist. I will show, however, that this is not necessarily the case.

Let us consider again the entities composed by members of S_1–S_n. In the context of the Lewisian theory, we can admit they are all organisms and call them O_1–O_n. If we interpret the concepts of identity and non-identity in the standard, strict way, O_1–O_n will all be different from each other. Strict identity is reserved for the relation of an object to itself. Any objects that do not completely overlap are non-identical. And since none of O_1–O_n completely overlap, they are non-identical, that is, different. Lewis, however, suggests a different interpretation, one that is more in tune with common sense and ordinary language. He accepts that the concept of strict identity applies to objects that completely overlap, but reserves the concept of strict non-identity for cases of objects which do not overlap at all, such as my computer and the Eiffel Tower. But in between these two extremes, there is a spectrum of objects that overlap to different degrees, thus falling under the concept of partial identity. At one end there are cases like that of Siamese twins connected only by a finger, and at the other there are objects that share almost every part, such as our organisms O_1–O_n. According to Lewis, objects in this spectrum are *partially identical*, and objects with very extensive overlap are *almost identical* (Lewis 1999, 178).

Just as the ordinary notion of identity differs from the concept of strict identity, so does the concept of counting. Strictly speaking, we count according to identity interpreted the standard way. If we do so, the sentence "Where I am, there is just one organism" will be false, because it is not true that O_1–O_n completely overlap. But in ordinary circumstances, says Lewis, we sometimes count according to relations other than identity (Lewis 1999, 175), and there is no reason why we could not use the concept of partial identity for counting, especially in cases where there is very extensive overlap. In such cases we can say the objects are almost identical. As a result, the above sentence will be almost true and by a blameless approximation we may say that where I am, there is just one organism. For most contexts, this will be true enough, according to Lewis.

6. The cheating revealed

Both the supervaluationist solution and the solution by partial identity enable the animalist to retain the premise of his argument against the psychological theory of personal identity: where I am, there is just one organism (one person, one thinker, etc.). On the supervaluationist solution, although the term "organism R" is vague and there are a number of precisifications that could be suitable as the denotatum of the term, on any such precisification it will be true that exactly one set contains all and only those particles that compose R. This gives the status of supertruth to the animalist premise. Using partial identity, the animalist may say that although strictly speaking there are many organisms where I am, since they overlap to such a great extent we may say they are almost identical, which is good enough for most contexts.

However, I have already indicated that these two solutions amount to a sort of cheating. Opponents will be quick to point out that the real issue is not what we may permissibly say on most occasions, but what is actually the case. And no matter what we say on these accounts, we still have to face the facts.

The supervaluationist solution fails, its opponents might say, because it merely carefully conceals the fact that the precisifications are all material entities, that they are all extremely similar, and that, as a result, they are all equally well suited to be the organism. The fact that we can speak *as if* there were just one organism, because nothing turns on which of these entities we pick as the denotatum of the term, cannot hide the simple fact that all these very similar entities exist and that we cannot provide a plausible selection principle for choosing between them. As a result, we must accept the fact that where I am, there are millions of other thinking and intelligent organisms, and I cannot point to a single feature that makes one rather than any other one me.

Similarly, the solution based on partial identity can easily be discredited, because it is simply a form of pretence. Certainly, we can pretend that there is just one organism where I am, because all of the organisms that exist there are so extremely similar that we can hardly tell them apart. But this is just a *façon de parler*. There are all sorts of ways of speaking, but when hard-pressed we would have to admit that even though the organisms are *almost* identical, they are not in fact *identical*.

7. Facing the facts

Suppose, then, that we admit that we have been cheating, and we face the facts. There are two questions that need to be answered. 1) Do the facts pose any special problems for the animalist as opposed to any reasonable person? 2) Does the admission enable the Lockean to score any points over the animalist? To both of these questions, my answer is no.

Regarding the first problem, we need to remember that the animalist takes the claim "Where I am, there is just one organism" to be a pretheoretical belief, an intuition, that most reasonable people normally accept. It is in all our interests to find a solution to the Problem of the Many, because it challenges this very intuition. The two suggested solutions attempt to save the intuition by saying, in their own ways, that in spite of the actual plurality, the intuition is still, in a sense, correct. If the solutions are deemed unacceptable, then we are all in trouble, for we are all deeply mistaken about how many objects there are. But the animalist, who says that where I am there is just one organism, is not in any deeper trouble than the cabinetmaker who says that she is working on a chest of drawers or the zookeeper who says that she is feeding an elephant. We all want there to be exactly as many things as we believe there are, and the supposition that there are a lot more than we think is disturbing to all of us. To put it another way, the animalist has the same beliefs about the number of animals in the world as ordinary people do. And whatever explication of those beliefs we must accept in order to account for vagueness, the animalist will be happy to accept, too.

But does the admission of plurality of organisms not give the Lockean the right to accept her preferred picture of personal identity? Does it not allow her to claim that where I am, there are two entities, an animal and a person? Could the Lockean not reason as follows?

> If the animalist can accept that where there seems to be one organism, there are, in fact, a great number of them, differing only by a single particle, then the Lockean can accept that where I am, there are two material entities, which do not differ by a single particle. After all, one particle surely cannot make any difference.

Well, it seems that she cannot, because there is a substantial difference between the plurality that the animalist accepts and the plurality that the Lockean defends. Remember that one premise of the Problem of the Many is that all of these alternative objects are so extremely similar that it would be completely unjustified to say one is an organism while the other is not. So they are all organisms. And whatever we normally wish to say about one of them will be true about all of the others. In particular, if one thinks, they all think, and if one has biological persistence conditions, they all do. This is what justifies us cheating, if anything does, and saying there is just one organism, one thinker, one biological continuer, etc.

The Lockean, in contrast, does not want to say that the animal and the person that it shares matter with are such that whatever one says about one of them will also be true of the other, despite the fact that they completely overlap. Some Lockeans say that the animal does not think (Shoemaker 1999), others claim it thinks only derivatively (Baker 2008), but all Lockeans say that the person cannot permanently cease to be conscious, whereas the animal can. So even if we tolerate cheating on the part of the animalist, we still have no justification for tolerating the idea that there can be two completely overlapping numerically distinct objects with different persistence conditions. No solution to the Problem of the Many that is available to the animalist licenses this Lockean claim.

Let's look at the dialectic of the dispute from the perspective of the two mysteries stated above. The animalist claims that the Lockean has to explain the metaphysical and the epistemic mystery. These mysteries arise, it will be remembered, because the Lockean claims that where I am, there are two material entities which completely overlap. Confronted with the vagueness of organisms, the animalist admits that where I am, there are multiple organisms which overlap not completely, but almost completely. Could the Lockean now say that the animalist has to explain the two mysteries, too?

I don't think so. The metaphysical mystery is a mystery about how two completely overlapping entities might differ in their persistence conditions. But the almost completely overlapping organisms do not differ in their persistence conditions! After all, they are all organisms, and they can all survive whatever an organism can survive. In particular, they can all survive the permanent loss of consciousness. The metaphysical problem is only a problem for the Lockean, because she claims that one of the overlapping entities cannot survive what the other can.

What about the epistemic mystery? Does the animalist not have to explain which of the almost completely overlapping organisms I am? Here the answer is trickier. Notice, however, that the epistemic question is troubling only if the answer has practical consequences. In the Lockean framework, the consequences are important. I need to know whether I am the animal or the person, because the answer entails what I can survive and whether it would be rational for me to visit Alcor, for instance. But in the case of the multiplicity of organisms the answer will have no practical implications, as we have seen. All of the organisms can pretty much survive the same things and get killed by the same things. So even if where I am there are millions of other organisms and I cannot tell which of them I really am, I am still confident that I cannot teleport myself and that developing an autoimmune disease may kill me even if Alcor intervenes at the last minute. So it is a mystery which of these multiple organisms I am, but it is one that is much easier to live (and die) with.[11] And let's not forget that this is a mystery that most of us have to solve, as we share the animalist intuition about the number of organisms.

Not everyone will be persuaded by this argument. It may be pointed out that I have ignored the real challenge that the problem of the many poses to the animalist and, instead, shifted attention to its practical implications.

[11] An anonymous reviewer has pointed out that the epistemic mystery dissolves on the semantic theory of vagueness, because the term 'I' will be vague in much the same way that the term 'organism' is. Thus, there is no answer to the question of which precise object I am, because 'I' does not refer determinately to any of them. I appreciate this comment.

But the problem is inherently metaphysical and epistemic, and it is in this light that it must be addressed.[12]

Two things can be said in response. First, it may be said that there is some value in purely metaphysical and epistemic arguments, but if nothing in practice turns on these arguments, the value is relatively low. Applying this line of reasoning, which is deeply rooted in the American Pragmatist tradition, we may say that the Lockean position is not disquieting merely because it prevents us from truly knowing whether we are persons or animals, but primarily because such knowledge is necessary for many of our practical interactions. The problem of vagueness of composition that the animalist must face, however, is merely theoretical. Perhaps it is a genuine epistemic issue that I am not able to tell which of the equally well-suited aggregates of matter I am. But nothing else hangs on it. Whether or not I am this or that aggregate of particles, I will be able to do the same things, survive the same changes and have the same mental capacities.

Secondly, the underlying assumption of the presented defence of animalism is the same as that which underlies the whole project of linguistic solutions to vagueness in general. These solutions also emphasize the practical aspects of the issue. The sceptic might object: ‘Look, perhaps we can talk *as if* there were just one cloud in the sky, but the real issue is a metaphysical one, not linguistic—there are many of them and we have no reason for preferring one of them. So, we cannot refer to ‘the cloud’ and make any statements about it due to referential failure.’ But the whole supervaluationist project is based on the idea that the practical issues trump the metaphysical ones. We do succeed in referring to the cloud and saying true things about it in spite of the fact that, metaphysically speaking, there are other equally good candidates. The important thing is that since most of what we say will be true regardless of which candidate is the right one, it might hurt that we cannot solve the metaphysical problem, but not very much.

This brings me to a final point related to the dialectic of the dispute. It has been noted that vagueness affects all composite material objects. The Lockean also believes in such objects. In fact, she believes in more composite

[12] I would like to thank another anonymous reviewer for this objection.

material objects than the animalist, for in her ontology there are persons *in addition to* animals, and both of these are material and composite. And if animals face the Problem of the Many, surely persons do as well. As a result, the Lockean is bound to believe that where I am, there are millions of almost completely overlapping organisms and millions of almost completely overlapping persons, each of which completely overlaps with one of the organisms, but is not identical to it and differs from it in persistence conditions. The Lockean seems to have a lot more to explain than the animalist. It is not that the Lockean cannot employ supervaluationism or partial identity to account for the vagueness of persons or animals. She surely can. But that is not the main problem the Lockean is facing. The main problem is how to account for their relationship: how to simultaneously maintain that persons are material entities that coincide with animals and that persons and animals are numerically distinct and have different persistence conditions; how to maintain that the animal is distinct from the person when it apparently has the mental properties sufficient for personhood. These are not problems induced by vagueness, so the solutions to the problem of vagueness are of no use there.

8. Conclusion

There are currently no generally accepted solutions to the Problem of the Many. Each solution has implications that clash with some of our intuitions. Solutions based on ontic vagueness or on brutal composition entail or are at least consistent with the idea that where I am, there is literally one material object. The solutions I have defended in this paper take a less direct approach, but still enable the animalist at least *to say* that there is one material object. All of these strategies are respectable, even if they have their critics. My goal has been to show that (a) the animalist is free to adopt the linguistic solutions to the problem, (b) this gives the Lockean no tools for defending her picture of personal identity, and (c) the situation for the Lockean is far more troubling.

Acknowledgements

I am indebted to Petr Dvořák for fruitful discussions of vagueness and related issues and to Eric Olson for helpful comments to a draft of this paper.

Funding

This work was supported by the Czech Science Foundation (GAČR), grant no. 17-12551S.

References

Baker, Lynne Rudder. 2000. *Persons and Bodies: A Constitution View*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139173124

Baker, Lynne Rudder. 2008. "Response to Eric Olson." *Abstracta* (special issue I): 43–45.

DeGrazia, David. 2005. *Human Identity and Bioethics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511614484

deRosset, Louis. 2011. "What is the Grounding Problem?" *Philosophical Studies* 156 (2): 173–197. https://doi.org/10.1007/s11098-010-9590-4

Geach, Peter T. 1980. *Reference and Generality*, 3rd edition. Ithaca: Cornell University Press.

Johnston, Mark. 1987. "Human Beings." *Journal of Philosophy* 84 (2), 59–83. https://doi.org/10.2307/2026626

Johnston, Mark. 2007. "'Human Beings' Revisited: My Body is not an Animal." In *Oxford Studies in Metaphysics*, vol. 3, edited by Dean Zimmerman, 33–74. Oxford: Oxford University Press.

Lewis, David. 1976. "Survival and Identity." In *The Identities of Persons*, edited by Amélie Oksenberg Rorty, 17–40. Berkeley: University of California Press.

Lewis, David. 1999. "Many, but Almost One." In *Papers in Metaphysics and Epistemology*, 164–182. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511625343.010

Markosian, Ned. 1998. "Brutal Composition." *Philosophical Studies* 92 (3): 211–249. https://doi.org/10.1023/A:1004267523392

McDowell, John. 1997. "Reductionism and the First Person." In *Reading Parfit*, edited by Jonathan Dancy, 230–250. Oxford: Blackwell.

McGee, Vann, and Brian P. McLaughlin, B. 2000. "The Lessons of the Many." *Philosophical Topics* 28 (1): 129–151. https://doi.org/10.5840/philtopics200028120

McKinnon, Neil. 2002. “Supervaluations and the Problem of the Many.” *Philosophical Quarterly* 52 (208): 320–339. https://doi.org/10.1111/1467-9213.00271

Olson, Eric T. 1997. *The Human Animal. Personal Identity without Psychology.* Oxford, New York: Oxford University Press. https://doi.org/10.1093/0195134230.001.0001

Olson, Eric T. 2007. *What Are We? A Study in Personal Ontology.* New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195176421.001.0001

Olson, Eric T. 2008. “Replies.” *Abstracta* (special issue I): 32–42.

Olson, Eric T. 2015. “What Does It Mean to Say that We Are Animals?” *Journal of Consciousness Studies* 22 (11–12): 84–107.

Parfit, Derek. 1984. *Reasons and Persons.* Oxford: Clarendon Press. https://doi.org/10.1093/019824908X.001.0001

Shoemaker, Sydney. 1984. “Personal Identity: A Materialist’s Account.” In *Personal Identity*, edited by Sydney Shoemaker and Richard Swinburne, 67–132. Oxford: Blackwell.

Shoemaker, Sydney. 1999. “Self, Body, and Coincidence.” *Proceedings of the Aristotelian Society* 73 (1): 287–306. https://doi.org/10.1111/1467-8349.00059

Snowdon, Paul F. 2014. *Persons, Animals, Ourselves.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198719618.001.0001

Unger, Peter. 1980. “The Problem of the Many.” *Midwest Studies in Philosophy* 5 (1): 411–467. https://doi.org/10.1111/j.1475-4975.1980.tb00416.x

van Inwagen, Peter. 1990. *Material Beings.* Ithaca: Cornell University Press.

Zimmerman, Dean. 2008. “Problems for Animalism.” *Abstracta* (special issue I): 23–31.

Organon F 26 (2) 2019: 228–248
https://doi.org/10.31577/orgf.2019.26203

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

RESEARCH ARTICLE

The Significance of the Relationship between Main Effects and Side Effects for Understanding the Knobe Effect

Andrzej Waleszczyński*[a] – Michał Obidziński*[b] – Julia Rejewska*[c]

Received: 26 March 2019 / Accepted: 19 October 2019

Abstract: The characteristic asymmetry in ascribing intentionality, known as the Knobe effect, is widely thought to result from the moral evaluation of the side effect. Existing research has focused mostly on elucidating the ordinary meaning of the notion of intentionality, while less effort has been devoted to the moral conditions associated with the analyzed scenarios. The current analysis of the moral properties of the main and side effects, as well as of the moral evaluations of the relationship between them, sheds new light on the influence of moral considerations on the attribution of intentionality in the Knobe effect. The moral evaluation of the relationship between the main and side effects is significant in that under certain circumstances it cancels asymmetry in intentionality ascription.

Keywords: Asymmetry; intentional action; intentionality; Knobe effect; moral evaluation; moral properties; side effect.

* Cardinal Stefan Wyszyński University

[a] ✐ Corresponding author. Institute of Philosophy, Faculty of Christian Philosophy, Cardinal Stefan Wyszyński University, Wóycickiego 1/3, 01-938 Warsaw, Poland
✉ a.waleszczynski@uksw.edu.pl

[b] ✉ m.m.obidzinski@gmail.com

[c] ✉ julia.rejewska@gmail.com

1. Introduction

In 2003, Joshua Knobe conducted an experiment on the tendency to ascribe intentionality to actions. Currently, it is referred to in literature as the Knobe effect, according to which people have a tendency to ascribe intentionality in cases of negative, but not positive, side effects (Knobe 2003a). An understanding of the nature of moral discernment and its characteristics plays an essential role in elucidating the influence of moral considerations on the ordinary concept of intentionality. In recent years, this issue has drawn much interest, with numerous empirical studies striving to explain the observed effect (Knobe 2003a, 2003b, 2006; Nadelhoffer 2004a, 2004b, 2006; Wright and Bengson 2009; Holton 2010; Sripada 2010, 2012; Sripada and Konrath 2011; Hindriks, Douven, and Singmann 2016).

It appears that scenarios patterned after Knobe's structure of stories contain other components subject to moral evaluation in addition to the side effect, i.e. the moral value of the main effect or the moral value of the relationship between the main effect and the side effect. Furthermore, whereas the explanations of the Knobe effect offered to date have predominantly focused on the moral evaluation of the side effect, these other components may carry differential moral properties in different scenarios. Given the above theoretical premises, the central objective of the present paper is to examine the contribution of other moral considerations, such as moral evaluations of the main effect or of the relationship between the main and side effects, to ascribing intentionality in side-effect cases. Of interest here, is whether evaluations of other scenario components significantly affect the aforementioned asymmetry in ascribing intentionality. For this reason, in this paper, we are interested in whether the moral evaluations of main effect and side effect—and the relationship between them—significantly influence the attribution of intentionality to actions. We do not intend to try to explain the Knobe effect, but to examine how the moral evaluation of effects impacts the ascription of intentionality to the side effect.

2. Intentional action and the Knobe effect

In his widely discussed paper "Intentional Action and Side Effects in Ordinary Language" Knobe (2003a) Knobe presented an interesting experiment concerning ordinary intuitions associated with ascribing intentionality. He presented respondents with two scenarios which were structurally identical in terms of intentional behavior theory, the only difference being the moral value of the side effects of the agent's actions, which had not been taken into account in standard approaches. One scenario represented a "help" version with the side effect being positive, while the other one contained a "harm" version, with the side effect being negative.

The scenario with the "harm" version was as follows:

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.' The chairman of the board answered 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was harmed. (Knobe 2003a, 191)

And the one with the "help" version was as follows:

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also help the environment.' The chairman of the board answered 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was helped. (Knobe 2003a, 191)

The respondents were asked whether the chairman of the board intentionally harmed or helped the environment (depending on the version of the scenario). It was found that they were more likely to ascribe intentionality when the side effect was negative (82%) versus positive (23%). Since then, numerous studies and analyses have corroborated the observed asymmetry in ascribing intentionality, which has come to be known as the “Knobe effect” or the “side-effect effect.” Knobe’s results have been also replicated in other languages, e.g., Hindi (Knobe and Burra 2006), German (Dalbauer and Hergovich 2013), and Polish (Kuś and Maćkiewicz 2016; Waleszczyński, Obidziński and Rejewska 2018), which indicates that the Knobe effect is culture- and language-independent, and as such may be successfully studied in the Polish language.

Knobe’s research was focused on the issue of intentionality. According to the *Simple View* of intentional action (SV) (Adams 1986; McCann 1987), if the agent does not intend to cause a certain effect, then she cannot bring it about intentionally. Following this line of thinking, it would be erroneous to ascribe intentionality to the side effects described in either scenarios, in which the chairman of the board makes an uncoerced decision to implement a new corporate program designed to cause a positive effect A. Thus, achieving A is clearly the chairman’s objective. At the same time, the chairman has been informed (he predicts) that the initiation of the new program will also result in an additional side effect B. The chairman states that he is solely interested in achieving A and is completely indifferent to B. In other words, the chairman indicates that B is not his intention. In the two scenarios, the variable is the moral value of B, which gives rise to asymmetry in ascribing intentionality to actions leading to B. Within the SV framework, the asymmetry would be explained as erroneous attributions in the “harm” scenario. However, the situation is more complex. If the moral evaluation of the side effect is taken to influence intentionality ascriptions, it must be recognized that moral evaluation is equally applicable to the main effect. Under the circumstances, the moral evaluation of the side effect may be affected by that of the main effect, and the resulting relationship between the moral evaluations of the two effects may bear on perceptions of the side effect. It should also be borne in mind that moral evaluation is not an ordinary instance of weighing costs and benefits (Mallon 2008). Therefore, it should be examined whether a change in the nature of the relationship between the two effects may alter the ascription of intentionality in side-effect cases.

3. Explaining asymmetry in ascribing intentionality – discussion

Existing research on asymmetry ascription in side-effect cases has focused on several major aspects. Knobe (2004, 2006) explained his findings in terms of the moral evaluation of the side effect. In his opinion, people tend to ascribe intentionality when the side effect is bad, but not when it is good. Following Hindriks, this explanation shall be called the *Moral Valence Hypothesis* (MVH).

However, it should be remembered that in his seminal experiment, Knobe (2003a) formulated two questions for each scenario. One concerned the chairman's intention to cause the side effect, while the other one asked respondents how much blame (in the "harm" version) or praise (in the "help" version) the chairman deserved for bringing it about. Knobe's results showed a correlation between attributing blame and intentionality. Analysis of these results has revealed yet another asymmetry, termed the *Praise–Blame Asymmetry* (Hindriks 2008, 630), which has given rise to a new approach to the Knobe effect. Blame attribution in this context has been explored in great detail by Hindriks et al. (2016), who have reported that intentionality ascriptions depend not so much on the *Praise–Blame Asymmetry*, as on the degree of attributed blame. However, this explanation does not hold in light of Knobe and Mendlow's study (Knobe and Mendlow 2004) which has revealed an asymmetry in ascribing intentionality in the absence of a tendency to attribute blame. Moreover, there also exist situations in which intentionality is not ascribed even though the side effect is negative (bad) and blame has been apportioned (Mele 2001; Nadelhoffer 2004). Indeed, such situations have refocused the researchers' attention on the concept of responsibility (Wright and Bengson 2009; Hindriks 2011). The observed correlation between ascribing intentionality and responsibility seems to shed more light on the Knobe effect than explanations based on the concept of blame as the former makes it possible to interpret situations in which responsibility is attributed in the absence of placing blame (Knobe and Mendlow 2004; Wright and Bengson 2009).

The above explanations of the Knobe effect and their underlying hypotheses refer to other concepts (blame, responsibility) and the related moral evaluations. A more autonomous explanation of the observed asymmetry is offered by Richard Holton (2010), who draws on the idea of a norm, and in particular norm violation. According to him, the asymmetry identified by Knobe results from the fact that people violate norms intentionally, while conforming to norms does not presuppose intentionality. In the "harm" version, the norm is violated, and consequently intentionality is ascribed, but in the "help" version the norm is observed, which is naturally interpreted as an instance of non-intentional behavior. From a philosophical perspective, the idea of a norm is similarly employed by Katarzyna Paprzycka (2014, 2015), who combines an orthodox theory of intentional action with a normative account of intentional omission. According to Paprzycka, the "harm" scenario entails an intentional omission to follow a norm. Therefore, the chairman's stated intention to achieve only the main effect does not prevent an ascription of intentional omission to observe a norm—the prerequisite for such an ascription is knowledge of the norm rather than an intention to violate it. At the same time, Paprzycka (2016) aptly observes that Holton's hypothesis about intentional norm violation (presupposing intention), presupposes intentional omission of a norm (presupposing knowledge). The main difficulty is that it is not known to what norm (if any) the respondents refer. If one assumed, as e.g., Shaun Nichols and Joseph Ulatowski (2007), that the tendency to asymmetrically ascribe intentionality in side-effect cases forms a stable pattern, the problem would only be exacerbated. This would imply that if one altered the content, but not the structure, of the scenario, then the violated norm would change as well. That would in turn mean that the respondents, in a predictable manner, each time refer to a violated norm which is different in each scenario. In other words, one would have to assume that in all experiments using Knobe's scenario structure the attitude of the respondents to the violated norm is predictable.

Finally, in F. Hindriks's *Normative Reason Hypothesis* (Hindriks 2008, 2011, 2014; Hindriks et al. 2016), the Knobe effect is explained by the agent's gradable indifference towards the side effect he has caused. According to Hindriks, in Knobe's scenario respondents perceive a certain obligation of the chairman to care about the consequences of his actions. In other words, Hindriks suggests that the chairman ignores a valid normative reason by expressing indifference. In ordinary speech, indifference is a propositional attitude sometimes interpreted in a categorical way and sometimes in a graded way. Complete indifference would be perceived as an attitude of neutrality, with maximum caring being its polar opposite (Hindriks et al. 2016, 215–16). Hindriks treats people's assessment of the chairman's indifference as a factor affecting the degree of intentionality ascribed to him, as indicated by prior research (Mele and Cushman 2007; Phelan and Sarkissian 2008; Guglielmo and Malle 2010). The higher the chairman's indifference towards respecting a normative reason, the higher the likelihood he will be attributed blame, and thus intentionality.

4. Examining the significance of the moral evaluations of the main and side effects for the Knobe effect

The previous explanations of the asymmetry appearing in the judgments regarding the intentionality of causing the side effect were focused, on the one hand, on the moral value of this effect (Knobe 2006), violation or omission of the recognized social norm (Holton 2010; Paprzycka) or the degree of indifference of the perpetrator to the resulting side effect (Hindriks 2014, 2016) and, on the other hand, on the dependencies between judgments on intentionality and the attribution of blame or responsibility (Wright and Bengson 2009; Hindriks 2011).

Studies carried out so far seem to unify a moral property, usually bringing it to one basic element. However, the philosophical analysis of moral problems takes into account more such properties. It takes into account, for example: intention, knowledge, consequences, circumstances and voluntary actions. In the case of an action that causes the predictable side effect, for the moral evaluation of the act, the relation between the moral value of the main effect and the moral value of the side effect is also important. If the relation of the main effect to the side effect is important for the moral evaluation, it may also be important for formulating the judgments of the intentionality of causing a side effect. To this end, we have formulated a main hypothesis, which states that the moral evaluation of the effects and relations between them significantly affects the attribution of intentionality to actions.

Therefore, in order to check this hypothesis, we assume that the asymmetry of the judgments regarding the intentionality of causing a side effect, which appears in the responses to questionnaires using the structure of the story scheme proposed by Knobe, is the model. In other words, in this article we will not be interested in either the common understanding of the concept of intentional action or identifying the conditions of its application. The purpose of our research is to check the influence of moral properties on the attribution of intentionality. In our experiments, the moral property will be the relationship that occurs between the moral evaluation of the main and side effects. The disappearance of asymmetry will testify to the verification of the adopted hypothesis and the significance of the studied moral properties for the emergence of the Knobe effect.

5. Experiment 1

The first goal of the presented experiments carried out, was to answer the following question: Does the relation between main or side effect have an influence on the Knobe's effect? The research hypothesis was that a modified relationship between the moral evaluations of the main and side effects (as compared to test (N1) with the "low-value main effect and high-value side effect" condition) would affect the ascription of intentionality. The second goal, was to investigate the properties of scenarios based on Knobe's structure that change the main effect to one that is highly valued—is its effect similar to the original one?

5.1. Method

In this study, scenarios in the Polish language[1] were administered to respondents in face-to-face settings. The experiment took place at different departments in Cardinal Stefan Wyszyński University in Warsaw. Students were assigned randomly to one of three experimental conditions in the "harm" or "help" version. The experiment and questions were presented in

[1] The trouble is that there is no clear correlate of the English adverb 'intentionally' in Polish. In our experiments we used the Polish adverb, "intencjonalnie."

a traditional—paper and pencil—fashion. Trail obtained 188 participants in this experiment (32 in each versions in the second condition, and 31 in each versions in two remaining conditions) (Obidziński and Waleszczyński 2019).

The first test, (N1), with the "low-value main effect and high-value side effect" condition employed Knobe's original scenarios (Knobe 2003a), as presented in the section, *Intentional action and the Knobe effect*. Respondents presented with the "harm" and "help" scenarios were asked "Did the chairman intentionally harm the environment?" and "Did the chairman intentionally help the environment?," respectively.

The second test (N2), with the "high-value main effect and medium-value side effect" condition involved scenarios based on Knobe's structure from test N1. However, the main effect was modified so that it would be objectively highly valued. It was decided that the development of a drug for a hitherto incurable type of cancer would meet this condition. The side effect was also conceived of as a disease to align it in the same category with the main effect. At the same time, it was assumed that pneumonia as a side effect would entail a relatively low moral evaluation. According to the research hypothesis, a change in the moral evaluations of the main and side effects would shift evaluations of the relationship between these effects, which would consequently impact the ascription of intentionality in side-effect cases.

The scenario with the "harm" version was as follows:

> The vice-president of an experimental oncological hospital went to the chairman of the board and said, "We are thinking of starting the production of a new medicine. It will help us cure patients of pancreatic cancer but it will also cause pneumonia." The chairman of the board answered, "I don't care at all about causing pneumonia. I just want to cure the patients of pancreatic cancer. Let's start the production of a new medicine." They started the production of a new medicine. Sure enough, the patients came down with pneumonia.

And the one with the "help" version was as follows:

> The vice-president of an experimental oncological hospital went to the chairman of the board and said, "We are thinking of starting the production of a new medicine. It will help us cure patients

> of pancreatic cancer but it will also cure them of pneumonia." The chairman of the board answered, "I don't care at all about curing pneumonia. I just want to cure patients of pancreatic cancer. Let's start the production of a new medicine." They started the production a new medicine. Sure enough, the patients were cured of pneumonia.

The respondents were asked the question "Did the chairman intentionally cure/cause pneumonia?," depending on the scenario version. The response scale was the same as for test N1.

A subsequent test (N3), with the "high-value main effect and low-value side effect" condition was designed in order to address this interpretational difficulty. The test employed Knobe's original scenarios, but with the main and side effects reversed. In this way, both the structure of the scenarios and the moral evaluations of the two effects remained unchanged. The only modification concerned the relationship between the effects. In this experiment, the research hypothesis was that a modified relationship between the moral valuations of the main and side effects (as compared to test (N1) with the "low-value main effect and high-value side effect" condition) would affect the ascription of intentionality.

The scenario with the "harm" version was as follows:

> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us help the environment, but it will also cause losses." The chairman of the board answered, "I don't care at all about causing losses. I just want to help the environment as much as I can. Let's start the new program." They started the new program. Sure enough, losses were caused.

And the one with the "help" version was as follows:

> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us help the environment, but it will also increase profits." The chairman of the board answered, "I don't care at all about increasing profits. I just want to help the environment as much as I can. Let's start the new program." They started the new program. Sure enough, profits were increased.

The respondents who were given the "harm" scenario were asked the question "Did the chairman intentionally cause losses?," and those who received the "help" scenario answered the question "Did the chairman intentionally increase profits?". The response scale was the same as for tests N1 and N2.

5.2. Results

First, the Shapiro–Wilk test was performed to check for normality of distribution, and it was found that none of the distributions met the normality criterion. Thus, analysis of differences between the study groups was conducted using the non-parametric Mann-Whitney *U* test. Due to the fact that all of its results were convergent with those of Student's *t*-test, the latter are presented in this paper.

The obtained data were analyzed using Student's *t*-test for independent samples to determine the presence or absence of ascription asymmetry and to establish whether the differences between the groups responding to different scenarios were statistically significant.

In the N1, the mean scores were in the "harm" version +1,36 (SD = 2.042) and in the "help" version, –1.16 (SD = 2.083) (F[2] = 0.235, p = .630; *t*(60) = 4.802, p < .001, and Cohen's $d_{unbiased}$ = 1.205). In turn, for N2, the mean scores were in the "harm" version +0.84 (SD = 2.05) and in the "help" version –0.78 (SD = 1.879) (F = 0.666, p = .406; t(62) = 3.306, p = .002, and Cohen's $d_{unbiased}$ = 0,817). In the N3 the mean scores were +1.65 (SD = 1.644) in the "harm" version and +0.39 (SD = 1.856) in the "help" version (F = 1.437, p = .235; t(60) = 2.825, p = .006, and Cohen's $d_{unbiased}$ = 0.709).

The result of *t*-test, for the differences between "harm" and "help" scores absolute values in N1 (M_{N1} = 4.387, SD_{N1} = 1.283) and N2 (M_{N2} = 3.687, SD_{N2} = 1.575) was not significant: F = 1.507, p = 0.264; t(61) = 1.93, p = 0.058. Finally, the results of *t*-tests for differences in mean scores in "help" story judgment, between all experimental conditions were tested. For groups N1 and N3: F = 0.709, p = .403; t(60) = –3.090, p = .003, and Cohen's $d_{unbiased}$ = 0.775. For groups N2 and N3: F= 0.030, p = .863; t(61) = –2.482, p = .016, and Cohen's $d_{unbiased}$ = 0.618.

[2] Fisher homogeneity test.

5.3. Discussion

The obtained results support the research hypothesis that the relationship between the moral evaluations of the main and side effects has a significant influence on the symmetry of intentionality ascriptions in side-effect cases. Moreover, there was significant difference between "help" score in groups (N1) with the "low-value main effect and high-value side effect" condition and (N3) with the "high-value main effect and low-value side effect" condition. The reversal of the main and side effects cancels the attribution asymmetry reported for the original scenario versions. Taking into account the fact that the main effect/side effect relation was the only thing that differentiates the two conditions it is very possible that the observed lack of Knobe effect is due to the given experimental manipulation. Moreover, the new scenario based on the Knobe scenario turn out similar effects to the standard Knobe's scenario, thus it was contradictory to our assumption. However, the probability value for this analysis is very close to the level of significance. Moreover, taking into account our assumptions, the one-tailed test result is significant (p = 0,029).

6. Experiment 2

In the second experiment we are investigating, whether the difference observed in the first experiment will appear once again in the more random sample—thus supporting the hypothesis. Moreover, once again, modification of scenario was tested.

6.1. Method

In this study scenarios in the Polish language were administered to respondents in face-to-face settings. The participants were random people encountered in the vicinity of the Warszawa Śródmieście and Główna Railway Stations as well as the Łódź Kaliska Railway Station. Participants were assigned randomly to one of the experimental conditions in the "harm" or "help" version. The experiment and questions were presented in the traditional—paper and pencil—fashion. Trail obtained 186 participants in this experiment (31 in each versions in all three conditions) (Obidziński and

Waleszczyński 2019). The used methodology was identical to the one used in experiment 1.

6.2. Results

Once again, the Shapiro–Wilk test was performed to check for normality of distribution, and it was found that none of the distributions met the normality criterion. Thus, analysis of differences between the study groups was conducted using the non-parametric Mann-Whitney *U* test. Again, because of convergent results of both tests, the student's *t*-test will be used.

In the N1, the "harm" and "help" versions, the mean scores were +1.94 (SD = 1.731) and –1.06 (SD = 2.265), respectively, on a seven-point scale ranging from +3 (definitely yes) to –3 (definitely no), with 0 designated as "hard to say." Statistical significance was confirmed by Student's *t*-test (F = 5.153, p = .027; t(56.130) = 5.860; p < .001) and Cohen's $d_{unbiased}$ (1.47). Thus, as expected, the study revealed a statistically significant Knobe effect. In turn, for N2, the mean score for the "harm" version was +0.36 (SD = 2.303), and that for the "help" version was –0.39 (SD = 2.14). While the results revealed asymmetry in ascribing intentionality, it was no longer statistically significant (F = 0.699, p = .406; t(60) = 1,314; p = .194). In the N3, the mean score for the "harm" version was +0.26 (SD = 1.57), and that for the "help" version was +0.16 (SD = 2.208). Thus, the results of the two scenarios were convergent and indicative of symmetry in ascribing intentionality (F = 7.988, p = .006; t(54.164) = 0.199; p = .843).

The result of *t*-test, for the differences between "harm" and "help" scores absolute values in N1 (M_{N1} = 4.613, SD_{N1} = 1.202) and N2 (M_{N2} = 3.774, SD_{N2} = 1.499) was significant: F = 1.555, p = 0.232; t(61) = 2.43, p = 0.018, $d_{unbiased}$ = 0.61. Finally, the results of *t*-tests for differences in mean scores in "help" story judgment, between all experimental conditions was tested. A significant difference was observed only for N1 and N3 conditions: F = 0.004, p = .953; t(60) = –2.158, p = .035, and Cohen's $d_{unbiased}$ = 0.541.

6.3. Discussion

The obtained results support the research hypothesis that the relationship between the moral evaluations of the main and side effects has a significant influence on the symmetry of intentionality ascriptions in side-effect cases. There were no significant differences between the "harm" and "help" versions in group (N3) with the "high-value main effect and low-value side effect" condition. Moreover, there was significant difference between "help" scores in N1 and N3 groups. The reversal of the main and side effects cancels the attribution asymmetry reported for the original scenario versions. Taking into account the fact that the main effect/side effect relation was the only thing that differentiates the two conditions it is very possible that the observed lack of Knobe effect is due to given experimental manipulation. Second, there was a significant difference between group (N1) with the "low-value main effect and high-value side effect" condition and group (N2) with the "high-value main effect and medium-value side effect" condition results. It supports our assumption that changing the main effect on one valued higher will affect the asymmetry.

7. General discussion

The point of reference for the present study was the Knobe effect, or asymmetry in ascribing intentionality in side-effect cases. The results of group (N2) with the "high-value main effect and medium-value side effect" condition indicate that intentionality ascriptions may be affected not only by the agent's indifference towards the consequences of his actions, but also by a change in the moral evaluations of the main and side effects. Test (N3) with the "high-value main effect and low-value side effect" condition has corroborated the influence of the examined moral properties on intentionality attributions and made it possible to elucidate their nature. It has been found that of greatest significance is the relationship between the moral valuations of the main and side effects. Indeed, this relationship is critical to asymmetry in intentionality ascriptions. The results of test (N3) with the "high-value main effect and low-value side effect" condition for the "help" version are significantly statistically different from those for test (N1) with the "low-value main effect and high-value side effect" condition. However, of particular importance is the fact that symmetry was obtained by a radical increase in intentionality ascriptions in the "help" version, which must be surprising from the SV perspective.

Previous efforts to explain the Knobe effect were more focused on intentionality ascriptions in the "harm" version, as those ascriptions appeared to be inconsistent with the SV. The experiments presented in this paper shed new light on prior studies exploring the notion of intentional action. The aforementioned findings from works analyzing blame apportioning seem deficient as the emergence of intentionality ascriptions in the "help" version would entail blame attribution, which is a contradiction in terms in light of the meaning of the notions of blame and morally positive effects. Therefore, the Knobe effect cannot be explained by blame apportioning, and in particular by the *Praise–Blame Asymmetry*, which only reveals an existing correlation emerging under certain specific circumstances. As regards Hindriks's *Normative Reason Hypothesis*, indifference towards the side effect should be acknowledged as a significant factor in ascribing intentionality, but it is nevertheless secondary to the relationship between the moral evaluations of the main and secondary effects. Already test (N2) with the "high-value main effect and medium-value side effect" condition showed that a change in those evaluations influenced the extent of ascribed intentionality with respect to test N1. It may be expected that variation in the degree of indifference may additionally modify intentionality attributions, but that factor is unlikely to be decisive in accounting for the observed attributional asymmetry. Indeed, it seems that Hindriks overestimated the role of indifference in explaining the Knobe effect. Also Holton's and Paprzycka's proposals do not seem to hold in light of the presented new experimental results. While their findings explain intentionality ascriptions in the "harm" version, both authors' hypotheses would be falsified if applied in the "help" version as causing a positive side effect could hardly be shown to violate any moral norms.

On the other hand, it should be noted that the presented evidence does not contradict Knobe's MVH. In explaining attribution asymmetry, Knobe proposed that it was influenced by the moral evaluation of the side effect, which is correct, but does not account for the other factors at play. While Knobe was right that moral considerations, and especially the moral evaluation of the side effect impact the ascription of intentionality in bringing it about, it has been found here that the influence of moral considerations and the moral evaluation of effects is more complex than previously thought.

The Knobe effect could be explained by a new hypothesis, proposed here as the *Wide Moral Valence Hypothesis*, according to which asymmetry in ascribing intentionality in side-effect cases is attributable to one main underlying cause, which is the moral evaluation of the relationship between the moral values associated with the main and side effects. The existence of this factor, that is, moral properties affecting the way people perceive complex situations, has been indicated by P. Egré and F. Cova (2015). They reported that the moral considerations associated with negatively valenced concepts, such as death, and positively valenced ones, such as survival, bear significantly on the way people think and perceive the world, and consequently, on the way they arrive at their evaluations. The mechanism used in Egré and Cova's study, that is, reversing the order of responses, did not affect the Knobe effect (Nichols and Ulatowski 2007). However, the present test (N3) with the "high-value main effect and low-value side effect" condition did produce results somewhat convergent with Egré and Cova's work in terms of altering the valence of evaluations. Analysis of Egré and Cova's findings in conjunction with the present evidence suggests that along with their positive and negative aspects, the effects of actions have additional attributes in the form of moral properties. If one rejects the hypothesis about the existence of the moral properties of effects, then in the "help" version of test (N3) with the "high-value main effect and low-value side effect" condition the relationship between the positive main effect and the positive side effect would remain identical to the analogous relationship from test (N1) with the "low-value main effect and high-value side effect" condition in terms of moral evaluation. If the ascription of intentionality were influenced solely by the positive dimension of effects, then the reversal (swapping) of the main and side effects should not significantly affect the respondents' ascriptions of intentionality to the agent causing the side effect. However, such a reversal did in fact have a significant impact on intentionality attributions. This means that the positive dimension of effects must also have some moral properties. Given that an analogous relationship exists between the positive and negative effects, it may be argued that the emergence of the Knobe effect depends on the relationship between the moral properties of the main and side effects.

Two questions remain open. One concerns the way in which the moral properties of effects are discerned, and the other one the way in which their significance for a given situation is determined, or "measured." It should be remembered that such discernment or "measurement" do not have to be made in a purely rational way and that they do not amount to a simple weighing of costs and benefits (Machery 2008; Mallon 2008). Therefore, it cannot be excluded that in situations where negative moral values are discerned, the process of making ordinary moral evaluations is governed by mechanisms that in some ways differ from those governing evaluations of situations characterized by positive moral values. These issues certainly require further study.

8. Summary

The objective of the present study was not so much to provide another explanation for the Knobe effect, as to test the hypothesis according to which moral evaluations of the main and side effects and the relationship between them significantly influence the attribution of intentionality to actions. Furthermore, acknowledging the crucial role of such moral evaluations, it seems reasonable to propose that certain situations are characterized by specific moral properties. It has been found that in cases of side effects the ascription of intentionality (which is distinct from passing a moral judgment) depends not only on whether the effects in question are positive or negative, but also on whether they are perceived to have positive or negative moral value. Of greatest importance is the relationship between the moral evaluations of the main and side effects. The underlying cause of this finding may be theoretically determined and identified as a factor that is crucial to human intuitions and judgments. This implies that the ordinary meaning of words is associated with certain moral properties that underpin moral evaluations. As it was shown in test (N3) with the "high-value main effect and low-value side effect" condition, those properties play a significant role in intuitions of non-ethical nature. The identification of such properties, which requires further study, constitutes a challenge to moral language and metaethical theories. It may well be that moral language goes beyond utterances containing terms such as duty, obligation, and responsibility, and categories such as good/bad and praise/blame, and that it encompasses a wide spectrum of utterances and words exhibiting certain moral properties. Using the language of psychology, one could argue for the existence of a mechanism of axiological attribution which would associate different states of affairs (situations, normative systems), actions, or even individuals, with specific moral properties, which may be additionally modified by other moral properties inherent in ordinary utterances.

In conclusion, it should be added that the results of the present experiments suggest that the Knobe effect is mostly attributable to the moral evaluation of the relationship between the moral properties of the main and side effects. Previous replications of Knobe's seminal experiment were successful because they held the moral evaluation of the relationship between the two effects constant, and so the ascription asymmetry was reproduced. However, the current experiments, and in particular test (N3) with the "high-value main effect and low-value side effect" condition, showed that a modification of the moral evaluation of the relationship between the main and side effects may cancel the Knobe effect.

References

Adams, Frederick. 1986. "Intention and Intentional Action: The Simple View." *Mind and Language* 1 (4): 281–301. https://doi.org/10.1111/j.1468-0017.1986.tb00327.x

Dalbauer, Nikolaus, and Andreas Hergovich. 2013. "Is What is Worse More Likely? The Probabilistic Explanation of the Epistemic Side-Effect Effect." *Review of Philosophy and Psychology* 4 (4): 639–57. https://doi.org/10.1007/s13164-013-0156-1

Egré, Paul, and Florian Cova. 2015. "Moral Asymmetries and the Semantics of 'Many.'" *Semantics and Pragmatics* 8 (13): 1–45. https://doi.org/10.3765/sp.8.13

Guglielmo, Steve, and Bertram F. Malle. 2010. "Can Unintended Side Effects be Intentional? Resolving a Controversy over Intentionality and Morality." *Personality and Social Psychology Bulletin* 36 (12): 1635–47. https://doi.org/10.1177/0146167210386733

Hindriks, Frank. 2008. "Intentional Action and the Praise Blame Asymmetry." *Philosophical Quarterly* 58 (233): 630–41. https://doi.org/10.1111/j.1467-9213.2007.551.x

Hindriks, Frank. 2011. "Control, Intentional Action, and Moral Responsibility." *Philosophical Psychology* 24 (6): 787–801. https://doi.org/10.1080/09515089.2011.562647

Hindriks, Frank. 2014. "Normativity in Action: How to Explain the Knobe Effect and Its Relatives." *Mind and Language* 29 (1): 51–72. https://doi.org/10.1111/mila.12041

Hindriks, Frank, Igor Douven, and Henrik Singmann. 2016. "A New Angle on the Knobe Effect: Intentionality Correlates with Blame, not with Praise." *Mind and Language* 31 (2): 204–20. https://doi.org/10.1111/mila.12101

Holton, Richard. 2010. "Norms and the Knobe Effect." *Analysis* 70 (3): 417–24. https://doi.org/10.1093/analys/anq037

Knobe, Joshua. 2003a. "Intentional Action and Side Effects in Ordinary Language." *Analysis* 63 (3): 190–94. https://doi.org/10.1093/analys/63.3.190

Knobe, Joshua. 2003b. "Intentional Action in Folk Psychology: An Experimental Investigation." *Philosophical Psychology* 16 (2): 309–24. https://doi.org/10.1080/09515080307771

Knobe, Joshua. 2004. "Folk Psychology and Folk Morality: Response to Critics." *Journal of Theoretical and Philosophical Psychology* 24 (2): 270–79. https://doi.org/10.1037/h0091248

Knobe, Joshua. 2006. "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology." *Philosophical Studies* 130 (2): 203–31. https://doi.org/10.1007/s11098-004-4510-0

Knobe, Joshua, and Arudra Burra. 2006. "The Folk Concepts of Intention and Intentional Action: A Cross-Cultural Study." *Journal of Cognition and Culture* 6 (1–2): 113–32. https://doi.org/10.1163/156853706776931222

Knobe, Joshua, and Gabriel Mendlow. 2004. "The Good, the Bad and the Blameworthy: Understanding the Role of Evaluative Reasoning in Folk Psychology." *Journal of Theoretical and Philosophical Psychology* 24 (2): 252–58. https://doi.org/10.1037/h0091246

Kuś, Katarzyna, and Bartosz Maćkiewicz. 2016. "Z rozmysłem, ale nie specjalnie. O językowej wrażliwości filozofii eksperymentalnej." *Filozofia Nauki* 24 (3): 73–102.

Machery, Edouard. 2008. "The Folk Concept of Intentional Action: Philosophical and Experimental Issues." *Mind and Language* 23 (2): 165–89. https://doi.org/10.1111/j.1468-0017.2007.00336.x

Mallon, Ron. 2008. "Knobe versus Machery: Testing the Trade-off Hypothesis." *Mind and Language* 23(2): 247–55. https://doi.org/10.1111/j.1468-0017.2007.00339.x

McCann, Hugh J. 1987. "Rationality and the Range of Intention." *Midwest Studies in Philosophy* 10 (1): 191–211. https://doi.org/10.1111/j.1475-4975.1987.tb00540.x

Mele, Alfred R. 2001. "Acting Intentionally: Probing Folk Notions." In *Intentions and Intentionality: Foundations of Social Cognition*, edited by Bertram Malle, L. J. Moses, and Dare Baldwin, 27–43. Cambridge: MIT Press.

Mele, Alfred R., and Fiery Cushman. 2007. "Intentional Action, Folk Judgments, and Stories: Sorting Things out." *Midwest Studies in Philosophy* 31 (1): 184–201. https://doi.org/10.1111/j.1475-4975.2007.00147.x

Nadelhoffer, Thomas. 2004a. "Blame, Badness, and Intentional Action: A Reply to Knobe and Mendlow." *Journal of Theoretical and Philosophical Psychology* 24 (2): 259–69. https://doi.org/10.1037/h0091247

Nadelhoffer, Thomas. 2004b. "On Praise, Side Effects, and Folk Ascriptions of Intentionality." *Journal of Theoretical and Philosophical Psychology* 24 (2): 196–213. https://doi.org/10.1037/h0091241

Nadelhoffer, Thomas. 2004c. "The Butler Problem Revisited." *Analysis* 643 (3): 277–84. https://doi.org/10.1111/j.0003-2638.2004.00497.x

Nadelhoffer, Thomas. 2006. "Bad Acts, Blameworthy Agents, and Intentional Actions: Some Problems for Juror Impartiality." *Philosophical Explorations* 9 (2): 203–19. https://doi.org/10.1080/13869790600641905

Nichols, Shaun, and Joseph Ulatowski. 2007. "Intuitions and Individual Differences: The Knobe Effect Revisited." *Mind and Language* 22 (4): 346–65. https://doi.org/10.1111/j.1468-0017.2007.00312.x

Obidziński, Michał, and Andrzej Waleszczyński. 2019. "The Significance of the Relationship between Main Effects and Side Effects for Understanding the Knobe Effect: Database." OSF. February 2. osf.io/ky3re.

Paprzycka, Katarzyna. 2014. "Rozwiązanie problemu Butlera i wyjaśnienie efektu Knobe'a." *Filozofia Nauki* 22 (2): 73–96.

Paprzycka, Katarzyna. 2015. "The Omissions Account of the Knobe Effect and the Asymmetry Challenge." *Mind and Language* 30 (5): 550–71. https://doi.org/10.1111/mila.12090

Paprzycka, Katarzyna. 2016. "Intention, Knowledge, and Disregard of Norms: The Ommisions Account and Holton's Account of the Assymetric Intentionality Attributions." In *Uncovering Facts and Values: Studies in Contemporary Epistemology and Political Philosophy*, edited by Adrian Kuźniar and Joanna Odrowąż-Sypniewska, 204–33. Leiden - Boston: Brill Rodopi. https://doi.org/10.1163/9789004312654_015

Phelan, Mark T., and Hagop Sarkissian. 2008. "The Folk Strike Back; Or, Why You Didn't Do It Intentionally, though It Was Bad and You Knew It." *Philosophical Studies* 138 (2): 291–98. https://doi.org/10.1007/s11098-006-9047-y

Sripada, Chandra Sekhar 2010. "The Deep Self Model and Asymmetries in Folk Judgments about Intentional Action." *Philosophical Studies* 151 (2): 159–76. https://doi.org/10.1007/s11098-009-9423-5

Sripada, Chandra Sekhar. 2012. "Mental State Attributions and the Side-Effect Effect." *Journal of Experimental Social Psychology* 48 (1): 232–38. https://doi.org/10.1016/j.jesp.2011.07.008

Sripada, Chandra Sekhar, and Sara Konrath. 2011. "Telling More Than We Can Know About Intentional Action." *Mind and Language* 26 (3): 353–80. https://doi.org/10.1111/j.1468-0017.2011.01421.x

Waleszczyński, Andrzej, and Michał Obidziński, and Julia Rejewska. 2018. "The Knobe Effect from the Perspective of Normative Orders." *Studia Humana* 7 (4): 9–15. https://doi.org/10.2478/sh-2018-0019

Wright, Jennifer C., and John Bengson. 2009. "Asymmetries in Folk Judgments of Responsibility and Intentional Action." *Mind and Language* 24 (1): 24–50. https://doi.org/10.1111/j.1468-0017.2008.01352.x

Organon F 26 (2) 2019: 249–264 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2019.26204 ISSN 1335-0668 (print)

RESEARCH ARTICLE

Knowledge after the End of Nature: A Critical Approach to Allen's Concept of Artifactuality

Sezen Bektaş*

Received: 9 July 2018 / Accepted: 5 November 2018

Abstract: Barry Allen's criticism of the traditional definition of knowledge seems to share a radical tone with Stephan Vogel's concerns about the customary representation of the causes that lie behind our current environmental problems. Both philosophers voice their complaints about the Cartesian picture of the world and dismiss the core idea behind the notorious duality embedded in that picture. What they propose instead is a monistic perspective positing an artifactual networking. In this paper, I will try to draw attention to certain weak aspects of Allen's refreshing description of knowledge as "superlative artifactual performance" and offer a way to improve that characterization via Vogel's notion "wildness". More specifically, I will propose a solution to the problems pertaining to the distinction between good and bad artifacts with respect to the epistemic criteria proposed by Allen, and claim that the temporal gap standing in between the expectations of a designer and the qualities of her design may contribute to our understanding of the nature of an artifact. I maintain that each creative attempt to know a given artifact is to be appreciated by recognizing its different uses. In doing so, I will also try to show why and how certain bad artifacts get their undesirable status because of leading up to techno-cultural stagnation.

Keywords: Allen; artifactuality; knowledge; use-value; Vogel; wildness.

* Middle East Technical University
✐ Philosophy Department, Middle East Technical University, Dumlupinar Bulvari, 1, 06800 Ankara, Turkey
✉ sezen.altug@metu.edu.tr

1. Introduction

In his *Thinking like a Mall: Environmental Philosophy after the End of Nature*, Stephen Vogel (2015) provides a discussion regarding the current status of nature, the environmental problems caused by human harm to nature and the ethical issues raised on this ground. The debate is of importance as it concerns not only the philosophers of environmental ethics but also any thinker entertaining similar questions about humans' responsibility in reflecting over and assessing the current situation of their environment. Epistemology, as well as environmental philosophy in general, is keen to point out the fallible presuppositions of humans resulted from their oblivion to the outside world and this situation creates a noteworthy partnership between these two branches of philosophy. Since the original claim pertaining to the end of nature belongs to the environmentalist group, my strategy in writing this paper first of all will be to adapt their evaluation about the hegemony of artifacts to the domain of epistemology. Secondly, I will point out how their attitude bears similarities to the observations made by Barry Allen with respect to the objects of knowledge—which he deems to be thoroughly artificial by definition. Finally, an attempt will be made to open a discursive channel by which these two philosophical matters can communicate.

At the very outset, one point should be clarified. I will not argue that epistemology and environmental philosophy are foreign to each other or pretend that I am the first one to broach the issue of collaboration. Rather, my principal aim is to contribute to an already established dialogue with a specific purpose. I will attempt to enhance Allen's four criteria [appropriateness to use, quality of design, fecundity, and symbiosis (Allen 2004, 72-74)], which are spelled out to assess whether our attempts to know are performed superlatively in the light of Vogel's views about the nature of artifacts. I regard the main purpose of this paper as constructing a common ground for defining any form of the knowledge that might be defended in the post-naturalist philosophy and proposing an understanding which may help us to separate the good artifacts that we generate by the act of knowing from bad ones. In this way, I hope to improve Allen's definition of "knowledge" as the superlative performance with artifacts (Allen 2004, 72) by means of Vogel's definition of "wildness"—which resides in the gap between the intention of a builder and the consequences of her artifact (Vogel 2015, 113). Since Allen's principles appear in certain cases to be less than decisive to label some instances of knowledge to be "bad artifacts" as seen in the examples of Auschwitz or atomic bombs, I am inclined to think that importing certain ideas from another trend of philosophy to fully develop his assessment might prove philosophically fruitful.

Thus, I will commence my treatment by analyzing Vogel's rejection of the Cartesian picture of nature, which is the duality between the human and the non-human worlds. Lying at the center of his critical view is the idea of defining nature through an exclusion of human existence. I will try to offer a detailed perspective on his suggestion about the re-establishment of the relation between humans and nature, and the reconstruction of their worlds on a common ground where nothing can escape being artificial. Secondly, I will provide a construal of his Heideggerian thesis about the end of nature–which is the logical and ontological impossibility of encountering an untouched landscape. Originally, the view that the nature has already ended by human destruction belongs to another environmental philosopher, Bill McKibben (1989). The genuine contribution of Vogel to his claim is that the end of nature is not a recent occurrence. Rather nature has *always* already ended (Vogel 2015, 25). I will scrutinize the epistemic consequences of such a judgment later in this paper. Thirdly, I will show how Vogel's and Allen's reflections about the current stage of the civilized world are alike in certain significant ways. Although they develop their ideas in different areas of philosophy, they both take "web of artifacts" as the launching point of their inquiry. I will devote more space to Allen's opinions in the pertinent section and will endeavor to elaborate his creative understanding of the act of knowing. With an aim to shed light on certain problematic aspects of his representation of knowledge and to offer a way to improve it, I will propose a solution supported by Vogel's Derridean concept of "wildness" which is characterized as a temporal gap between the intention of the builder and the resulted qualities of the artifact (Vogel 2015, 113). More broadly, I hope to strengthen the hand of a refreshing standpoint about the problem of knowledge and to contribute to the ever-lasting process of eliminating defects of a promising theory.

2. Is there anything left that is natural?

In *Thinking like a Mall*, Vogel challenges the validity of the well-established dualism between nature and humans in environmental philosophy, a strategic move aimed at rendering agents responsible for their damage to nature. He maintains that there is something unsatisfactory about the whole controversy about the nature-human tension as he thinks that the term 'nature' is too ambiguous to be a reference point for positing (indirectly) what is not natural. The term has multiple meanings, and so is too unstable to be the main basis of the whole debate about the sources of our environmental problems.

> Each attempt to define nature falls prey to counterexamples that lead the definer to complain "no, that's not what I meant," and then to redefine the term yet again, in an ongoing dialectic that leaves one wondering at the end whether any clear sense can be made of the term at all. (Vogel 2015, 9)

The difficulty in giving an analytic definition to the concept "nature" is just the tip of the iceberg. When nature is examined ontologically, another and a more significant issue arises, to wit, the double nature of the term where we seem to have incompatible readings. Vogel approaches this problem by formulating the relationship between humans and nature depending on two different modes of being or states of nature. In one state, nature ontologically excludes humans on account of their capability of producing something unnatural. Consequently, by definition the human world turns out to be unnatural. In the other state or mode, nature encompasses humans because of their subjection to similar processes in evolution with other living beings. Thus, the human world is characterized to be inseparable from nature by definition.

Vogel benefits from John Stuart Mill's distinction in his *Nature* (1998) to familiarize the reader with his own analysis about the "double nature" of the term 'nature'. As Mill argues, in its first sense, nature stands for "the entire system of things, with the aggregates of all their properties"; and in its second sense, it denotes "things as they would be, apart from human intervention" (Mill 1998, 64). Similarly, Vogel assigns a word to each sense and employs the name 'Nature' (with capitalization) for "the totality of physical world", while "the nonhuman world" is called 'nature' with lowercase (Vogel 2015, 13). In this division, on the one hand, humans are benevolently depicted as the part of "Nature"; on the other hand, they are pictured as beings endangering "nature" violently enough to bring it to a dramatic end. Evidently, a problem occurs at this picture. The former definition causes a blatantly incorrect characterization of humans because their actions are identified as incapable of harming nature. The latter definition of nature is also odd because it gives the impression of implausibly relinquishing the lawful control of "nature" over to humans and almost granting them the freedom of doing whatever they want in their own "un-natural" world. In a nutshell, it removes human culpability vis-à-vis affecting and transforming "nature". Due to the logically, ontologically and ethically untenable implications of the dualistic representation of nature and humans, Vogel defends a monist perspective.[1] He marks each and every thing as unnatural or artificial, and constructs all other arguments on this unity in negation. Our presence in nature has *always* already transformed it into a built one (Vogel 2015, 29-30). All we can observe and experience is the artificialized or "built" environment.

3. Is there anything left that is natural to know?

Vogel's claim about ending nature by artificializing it and being obliged to live in post-naturalist environment has roots in the ideas of Bill McKibben. As the latter writer puts it:

> When I say that we have ended nature, I don't mean, obviously, that natural processes have ceased—there is still sunshine and

[1] Latour defends a similar position about the relationship between humans and non-humans in his *Politics of Nature* (2004). Nature and society are characterized as "two houses of a single collective", and the public life is organized in their association or intersection. Similarly, he advises that ecology focus on this common world instead of solely dealing with nature. This differs from Vogel's view because we are still talking about the areas where members of these two houses do not interact. Let me take this opportunity to thank the referee of *Organon F* for pointing out the relevance of Latour's work to my paper.

> still wind, still growth, still decay. Photosynthesis continues, as does respiration. *But we have ended the thing that has, at least in modern times, defined nature for us—its separation from human society.* (McKibben 1989, 64; emphasis in original)

It must be noted here that while McKibben contends that the nature is ended, he does not maintain that the damage caused by humans on nature cannot be undone. In that sense, for instance, some technologies which are geared towards preventing pollution or stopping global warming can still be utilized. However, even though we, as humans, are able to restore nature perfectly, this act to turn something into its original state will be a human artifact, and so will become unnatural anyway. We are not able to intervene with or relate to nature without transforming it in human ways. Or to put it differently, we cannot escape artificializing nature as long as we act in it.

McKibben's illustration of the end of nature can be simplified via an analogy comparing human relationship with nature to the touch of Midas (Vogel 2015, 11). The human touch alters nature every time agents establish some relationship with it or even they direct their attention to it. For Vogel, the history of this transformative relationship between humans and nature is as ancient as the history of human beings. The phenomenon of ending nature is not a recent event as it is popularly believed. As he says, "human beings have always transformed the world they encounter, and they transform it *in* encountering it, a fact that might well be part of their 'nature'" (Vogel 2015, 25). This shows us that the search for lands which have not been touched by humans, or by Midas metaphorically, is a "fetish" practiced by the dreamers of *wilderness.*[2] The reality does not correspond to the dreamers' frozen image of nature. The nature is not a *nature morte* or a thing that we can fix in an immutable state. Thus, it is conceptually

[2] *Wilderness* is a term which means "a tract or region uncultivated and uninhabited by human beings" (Merriam-Webster). The term also denotes the slogan of an eco-friendly act in 1964, which carried out protests for protecting pristine areas and for letting nature be. Later, the term also designated a long debate in environmental philosophy. The proponents of the wilderness are mainly criticized by J. Baird Callicott who argues for the replacement of this idea with a more realistic and still objective norm of "biodiversity". My objection to the idea of wilderness is basically a reflection of his *A Critique of and An Alternative to the Wilderness Idea* (1994).

and ontologically impossible to confront a piece of the Earth and to declare that it is un-touched, wild. The radical outcome of this reasoning is to drop the concept of nature for good in environmental theory and find another, a less objective standard (Vogel 2015, 28). For Vogel, there is no need to abstain from accepting an anthropocentric norm for the ethical foundation because the real source of our current environmental problems is not more but, rather, *less emphasis* on humans. Therefore, we should "develop an environmental ethic, and an environmental philosophy, that take the environment (a word that simply means 'what surrounds us') to be the built one we actually inhabit" and we should not be concerned with nature at all (Vogel 2015, 30).

Similar to Vogel's critique of "nature" in environmental theory, Barry Allen questions the effectiveness of preserving "truth" in epistemology. He disputes the adequacy of representational theories of knowledge which divide knowledge into two as *knowing-that* and *knowing-how*, allowing only the former to enter the territory of genuine knowledge (Allen 2008, 35). In a nutshell, the knowledge of "how" is characterized via our talents and habits such as knowing how to swim or to withdraw money from an ATM machine. These non-representational, non-verbal forms of knowledge are not truth-assignable which means that they can be neither true nor false. Moreover, this is evidently the reason why they are not exactly the favorite subject matter of those philosophers who build a notable career around the notion of propositional truth. The received view has it that in a significant sense (propositional) knowledge involves truth-value attributions in representational contexts. According to this logocentric approach, knowledge is designated as a thing which has a "true" essence and the representations of knowledge are assessed depending on whether they bear this essence or not. In parallel with Vogel's rejection of Cartesian duality between humans and nature, Allen asserts that this dichotomy of knowledge is not helpful because it prevents us to appreciate the value in *know-how* and it forces us to acknowledge only one-sided knowledge acquisition (Allen 2008, 36).

Our adventures of inventing knowledge have an evolutionary story according to Allen's reading. Human journey to "know" in sophisticated cultural contexts through artifacts has been continuing for about 40,000 years. We have been eliminating the predictable, habitual or ordinary aspects of our activity as we progressively refined our ways of preferring and selecting. In this way, we have managed to connect to surrounding things with which we are constantly engaged and whose reality only matters—i.e. quite simply, artifacts. Our knowledge turned into "a cultivated capacity for eliciting, creating, and amplifying superlative performance in artifacts".

> It is important that the performance be superlative, meaning not literally or uniquely *the* best, but *of* the best, among the best, at that rank. Knowledge, like art, can be found only in the best examples. Only superior performance necessarily implies knowledge. (Allen 2004, 62)

Contrary to the classical view in epistemology which qualifies knowledge through its reliability, he contends that knowledge actually necessitates a more refined and originality-based standard (Allen 2004, 67). Since we also treat our habits as reliable, reliability cannot be an adequate test for our performances to know. In that sense, our performances involving knowledge must set their own principles in each and every instance, and they must be assessed without requiring an isomorphism. However, another question deserves our attention at this point. What should be our reference in appraising the worth of the thing to be known? According to Allen, our environment is "saturated with artifacts" and their quality may only be evaluated by those who can understand how they function (Allen 2004, 88). We are surrounded by a network of artifacts which "presuppose each other, produce each other, work with and upon each other, in a web of interdependence now practically coextensive with the global human ecology" (Allen 2004, 64). This complex structure whose components operate in a concordant and co-dependent manner is the very condition of our knowledge.

4. Knowing the artifact by recognizing its wildness

As stated in the previous section, viewed from Allen's epistemological perspective, knowledge determines its own standard of appraisal. It is important to notice that this style in epistemology does not necessarily entail that we cannot have some objective parameters to judge knowledge. Some criteria can still be formulated if we let knowledge to decide them with reference to its own "traditions" of accomplishment. Allen lists four dimensions on which we can confirm that a given performance as an artifact deserves to be called "superlative" or "good" in character, which are appropriateness to use, quality of design, fecundity, and symbiosis (Allen 2004, 72). First of all, an artifact can be evaluated in terms of being user-friendly or not. Its adequacy in performing the task that it is designed to do is a crucial specification. Secondly, the design of the artifact should be authentic. Being useful or being functional, on its own, is not enough to qualify an artifact as the superlative form of achievement. Thirdly, an artifact should be the source of productiveness and should inspire others within the related fields to innovate. The richness in content and diversity in the application area bring advantages to an artifact, and enable it to offer new opportunities for the use different than its originally defined function. Lastly, the value of the artifact is proportional to the complexity of the relationship that it establishes with other artifacts. If an artifact is successful in making mutually beneficial connections with its environment, it becomes an irreplaceable artifact. These qualifications bestow a special epistemological status to the objects of the world as to render that world an artifactual one.

Allen makes an open-ended list to exemplify the "good" artifacts, which satisfy all four criteria and manage to pass the test of excellence, and to distinguish them from the "bad" artifacts (Allen 2004, 73). The artifacts such as the writing, the sailing ship and the penicillin are designated as the accomplishments of knowledge, and so labeled as good. However, the guillotine, the atomic bomb and Auschwitz are characterized as bad artifacts due to the fact that they fail the test by violating at least one criterion of superlative artifactual performance. For instance, guillotine contradicts with at least two criteria, which are fecundity and symbiosis. The function of guillotine is limited to kill and it does benefit the person who uses it (the executioner) but not the one for whom it is used (the executed). Thus, the guillotine as an artifact is neither productive nor capable to establish mutually beneficial relationships. In a nutshell, it falls under the category of bad artifacts.

As Allen argues, evaluating an artifact sometimes becomes more complicated and requires a more detailed reasoning. The existence of such perplexing examples seems to cause a paradox, and so poses a threat for the integrity of the test. Auschwitz is one of these obscure artifacts (Allen 2004, 74). When we put aside all cruelty and malice in the concentration camps and focus only on the technique, expertise and engineering which turned out to be necessary for the mass destruction, we may claim that Auschwitz is a superlative artifactual performance. However, making such a reduction would be cold-blooded as well as illegitimate because of breaking the relation of an event with its historical context. Therefore, it would be more appropriate to approach such events found in our collective memory as a unified phenomenon without isolating it from the values that historically burden it positively or negatively. Allen prefers analyzing this artifact specifically at the onto-epistemological level instead of conducting a discussion that incorporates the moral and political aspects of Auschwitz as well. His final decision about the quality of Auschwitz is negative because he asserts that "the camp was knowledge against itself" (Allen 2004, 74). The power gained by knowledge was used for destroying the ground which the knowledge requires to retain its sense of accomplishment. In that sense, Auschwitz cannot be acknowledged as a superlative artifactual performance because it violates a very basic principle, i.e. self-preservation.

Although Allen speaks to our conscience while listing Auschwitz under the label of "bad artifacts" one may still doubt the coherence of the logic which leads him to make this judgment. A reasonable objection in this context may be formulated as follows: Any instance of knowledge can be used to destroy itself, i.e. its own modes of generation. If self-destruction is to be conceived as a breach within the maxims of superlativeness, none of the examples of knowledge, including good artifacts, are exempt from this perilous prospect. Furthermore, self-annihilation should be reckoned as a capacity which is not inherent to the artifact; rather, it shows itself through its use-value. Even though the artifact performs exceptionally and appears to turn itself into a unit of knowledge, this does not necessarily mean the end of the story with respect to its criteria of qualification. In spite of the fact that an artifact is reasonably situated within a web of artifactual items, its place (and, thus, artifactual "goodness") may actually vary considerably depending on its current functional characteristics. Each new experience of its utility contributes to the total value of the artifact and alters its depiction.

A performance or act of knowledge can be evaluated negatively or positively depending on the intentions of those who transform the nature of artifact in question through their utilization. As for the example of Auschwitz, the totality of expertise, procedural techniques, and clusters of information can obviously be alleged to count as “knowledge”. Nevertheless, the main reason why those factors should not be deemed sufficient for such labeling is actually not due to some property of the artifact. Allen’s own reasoning in this context (viz. that in case of Auschwitz the end product stands against itself) is inadequate to constitute a negative instance; rather, one must contend that *the use of* artifact leading particularly to Auschwitz (combining the knowledge collected from different fields within the performance of mass destruction) actually causes the issue here. In that sense, the quality negating the superlative form should *not* be regarded as inherent to the accomplishment of an artifact but, rather, *external* to it. The manifestation and embodiment of the artifact in the form of Auschwitz assigns a negatively loaded history to it. However, as this exemplification is contingent to the nature of the artifact, it does not inherently need to address the evilness entangled with its epistemic characteristic.

Hence, my critique of Allen’s evaluation of the example of Auschwitz posits a difference between the quality in the character of the artifact and the quality of the value in its use. These two senses are attributed to the artifact at different “stages”. While the former is ascribed *before* (*antecedent to*) the artifact’s expressing knowledge, the latter is assigned *after* (*posterior to*) the artifact’s being expressed superlatively. This temporal gap between the distinctive types of qualifying an artifact can be understood as a version of what Stephen Vogel denotes as *wildness.*[3] He originally defines this unbridgeable gap between the intention of the builder and the properties of what is built. My intention, in the context of the critique offered here, is to employ Vogel’s insight and try to gesture at a critical rift between

[3] This term ‘wildness’ should be distinguished from the term ‘wilderness’ (Vogel 2015, 111). The former concept views the environment as a dynamic entity and appreciates the unpredictability or the creativity in its restoration. The latter concept, however, values not the whole environment, but rather merely the natural one. It picturizes this specific part of the environment as something stable, and so is against any sort of restoration of it.

the *characteristic of the artifact* and *its use-value*. To elucidate the definition of the concept, we can refer to Vogel's own writings:

> There is a gap, in the construction of every artifact, between the intention with which the builder acts and the consequences of her acts, a gap that is ineliminable and indeed constitutive of what it is to construct something; and in this gap resides something like what I earlier called wildness. And that gap, as I have just been suggesting, is not only the one between what we intend in our actions and the unintended consequences those actions nonetheless inevitably bring about but rather, and perhaps more important, it is there between our actions and their *intended* consequences, too, arising even when the object produced seems to turn out in just the way we had planned. It is a *temporal* gap, what Derrida would call a deferral, for even the successful execution of a plan requires, indeed depends upon, *waiting* for something that goes beyond the planning and beyond even the acts that put the plan into motion. (Vogel 2015, 113; emphasis in original)

As Vogel clearly states, some traits of artifacts are neither intended nor anticipated by their designers. In such examples, the nature of the artifact *exceeds* the intention of its designer. This specific quality of the artifact seems to have obvious connotations of *creativity* (Vogel 2015, 105).[4] As a matter of fact, occasionally the consequences become a surprise for the user as well. After the artifact is introduced into the "market" or is made public, some secondary–non-constitutive–values can be bestowed upon it through its use. I think the quality of knowledge to destroy its own existence is such a value.

Vogel's concept "wildness" would not be entirely foreign to Allen. His claim about the *intransitive* character of expressing an artifact bears

[4] I think this sense of creativity is very similar to Allen's description of the accomplishment of knowledge (Allen 2004, 68). He emphasizes the role of elegancy and innovation in denoting something as knowledge. He also exemplifies his view with a reference to the use of a paper clip. When we use it as a device to hold the sheets of a notebook together, we cannot be said to be exercising the capacity of human knowledge in the most appropriate way. Only when we use it creatively, for instance using a paper clip as an antenna, we may speak of the knowledge.

certain resemblances to Vogel's notion of wildness. Allen maintains that the artifactual expressions are "impersonal" in the sense that they do not "express a physical state first arising in the maker's soul" (Allen 2008, 38). He names this feature of expression specific to artifacts "intransitive expression":

> That is what aesthetic theory calls *intransitive* expression. Expression is transitive when it refers to an object, typically an emotional state of the maker [...]. When it comes to artifacts expression becomes intransitive. Like an intransitive verb, an expressive artifact doesn't require an "object," that is, psychological state of the maker that it transitively expresses. Intransitive, artifactual expressiveness *begins with the work*. It depends on how artifact is assembled, how it looks, not what the maker feels. (Allen 2008, 39)

Allen's "psychological state of the maker" is similar to Vogel's "intention of the builder", and neither Allen nor Vogel treats it as the unique determinant of the character of an artifact. Consequently, there is a significant discursive ground shared by Vogel and Allen. In this context, let me suggest another aspect of the matter in order to facilitate assessing the quality of the artifacts–call it "*modified* wildness". By slightly differing from Vogel's definition, I describe *wildness* as the temporal gap between the original expression of the artifact and its modified expressions-in-use. I am inclined to take the values which are assigned to the artifact posterior to its expression as fully open-ended rather than determined. The values may change in accordance with how artifacts are put into use by people. The use of technical knowledge may give rise to catastrophic consequences as in the example of Auschwitz. Knowledge of chemistry may turn into a deadly weapon in its use for atomic bombs. In these examples, the user's expectation in the conversion of information more or less overlaps with the results. However, this may not always be the case and the outcomes of the modification may be too hard to estimate even for the user. For instance, the newly discovered radium element was declared a benign artifact in the 1910s and expanded its market during the following years, including cosmetics and food sectors. Further investigation proved that radioactive products involved considerable risks, and so the radium as a fecund artifact lost its attraction with regard to its performative competitiveness in that field. This shows us that an artifact always retains its propensity for change; so that the user who intends to modify it must anticipate some surprise by virtue of the alterations she affects. In that sense, each and every use can be regarded an attempt to know with the proviso that some of them deserve to be called "creative".

5. Conclusion

What I have tried to accomplish in this paper can be characterized as a supplement to Allen's theory of knowledge with an aim to elaborate the idea of artifactual networking in our techno-social world. In my opinion, Vogel's anthropocentric post-naturalist environmental theory prepares a fertile ground for a variation on Allen's definition of knowledge as humans' superlative performance with artifacts. I tend to think that Vogel fruitfully names a notion we do come across in the unorthodox view of Allen, to wit, *wildness*. This rift standing between the intentions of the person who builds the artifact and the potentiality that the artifact possesses at the end reveals a weak side of the standards proposed by Allen to define human knowledge. The separation of artifacts as good and bad according to Allen's criteria becomes blurry in cases where the knowledge gained through superlative performances is re-expressed regardless of the authentic nature of the artifact and the intentionality of the inventor. This observation does not presuppose that there is an essence defining each artifact or that the initiatives to alter it cause this gap. On the contrary, this gap is inevitable due to the temporal difference between the conditions shaping the objectives of the designer and the circumstances defining the product. The resultant picture leads us to the following thesis: The *wildness* gains its full meaning and significance in the fact that the artifact unavoidably gains new qualities through its uses. Therefore, the richness in the expressions of the artifact increases in proportion to the diversity in its use. If we believe in the merits of Occam's razor on this matter, we cannot in my opinion regard every single use transforming an artifact as the postulation of a brand new artifact. Rather, the emphasis must be placed on the abundance in the ways of attaining the superlative performance, and the instances of alterations in usage should be collected under a common title—which defines the dynamic nature of the artifact comprehensively through developing a projection on both its anterior and possibly posterior expressions.

Lastly, I would like to consider and respond to a prima facie strong potential objection. One may question why we should wait for the "posterior" effects of Auschwitz to surface in order to declare that it is in fact a bad artifact. Given the atrocious outcomes of concentration camps, it may seem obviously misguided to suggest that one has to wait to see the results of their utilization in order to reach a decision about their quality as artifacts. Furthermore, one may justifiably argue that even the "antecedent" nature of such an artifact should suffice to label it as having extremely poor quality within the boundaries of the notion of "being superlative". My concise response to this objection is that it seems both logically sound and politically correct to insist that the hypothetical "value" of the knowledge of Auschwitz can never be intrinsic but rather is always instrumental. When it is under consideration as a candidate of "knowledge" in the sense explained in this paper, the value it may be alleged to possess has never been inherent in the material elements of Auschwitz. I believe that this historical case is inevitably to be catalogued as good or bad *for* what it was meant to lead to—i.e. massacring of millions of people. In that context, it was doomed to fail as superlative performance as it must be qualified over what and how it was *used for*.

The crucial point is that Auschwitz's antecedent and posterior characteristics are inseparable because its *raison d'être* precisely coincides with its use. Auschwitz was *inter alia* a historical phenomenon which yielded a form of knowledge which was compatible with a certain use, to wit, mass destruction. The antecedent and posterior qualities of Auschwitz are equated at the stage of *utilization* in such a manner that the product halts at a level of techno-cultural stagnation. In a nutshell, it petrifies and taints its "value". This is to be contrasted with the usage of artifacts such as penicillin or computer. In those "good" examples, there is still a risk of being abused through bad uses in the future. They may be manipulated to turn out to be biological weapons or Terminators. However, given that in case of the *original* emergence of items like penicillin the associated practices define a field of "superlative performativity" in the sense of Allen's criteria for proper knowledge, the badly transformed artifacts clearly fail vis-à-vis some of those (e.g. symbiosis). Each particular use of an artifact transforms the pertinent nexus encompassing other artifacts. Consequently, when we pass judgment on the knowledge-value of an artifact like Auschwitz, we cannot focus merely on the technical properties of such a construct (e.g. in terms of its material quality or efficiency) and talk about its adequacy within certain narrow operational parameters. The net upshot of these considerations is that Auschwitz proves to be a poor exemplification of "superlative artifactual performance" despite the fact that out of the ingenuity of some engineers, the whole project, hypothetically speaking, could possibly be made to "function better" in its presupposed purpose of mass destruction.

Acknowledgements

I am indebted to Murat Baç for his priceless comments and constructive criticisms on the earlier versions of this paper. I also thank a referee of this journal for their helpful suggestions.

References

Allen, Barry. 2004. *Knowledge and Civilization.* Boulder, CO: Westview Press.

Allen, Barry. 2008. *Artifice and Design.* London: Cornell University Press.

Callicott, John Baird. 1994. "A Critique of and an Alternative to the Wilderness Idea." *Wild Earth* 4 (4): 54–59.

Latour, Bruno. 2004. *Politics of Nature: How to Bring the Sciences into Democracy.* Translated by Catherine Porter. Massachusetts: Harvard University Press.

McKibben, Bill. 1989. *The End of Nature.* New York: Anchor Books.

Mill, John Stuart. 1998. "Nature." In *Three Essays on Religion: Nature, The Utility of Religion, Theism*, 3–65. New York: Prometheus Books.

Vogel, Stephen. 2015. *Thinking like a Mall: Environmental Philosophy after the End of Nature.* Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262029100.001.0001

Organon F 26 (2) 2019: 265–286
https://doi.org/10.31577/orgf.2019.26205

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

RESEARCH ARTICLE

Prospects for Experimental Philosophical Logic

Jeremiah Joven Joaquin*

Received: 24 April 2018 / Accepted: 19 November 2018

Abstract: This paper focuses on two interrelated issues about the prospects for research projects in experimental philosophical logic. The first issue is about the role that logic plays in such projects; the second involves the role that experimental results from the cognitive sciences play in them. I argue that some notion of logic plays a crucial role in these research projects, and, in turn, the results of these projects might inform substantive debates in the philosophy of logic.

Keywords: Applied logic; descriptive models; experimental philosophical logic; normative models; philosophy of logic; prescriptive models; psychology of reasoning; pure logic.

1. Introduction

Over the years, there has been a steady growth of research projects, which fall under what has been dubbed as *experimental philosophical logic* (Ripley 2016).[1] Like most sub-areas of philosophy that experienced the

[1] The following works might arguably be classified under this heading: (Alxatib and Pelletier 2011), (Bonini et al. 1999), (Brauner 2014), (Cobreros et al. 2012),

* De La Salle University
✐ Department of Philosophy, De La Salle University, 2401 Taft Avenue, 0922 Manila, Philippines
✉ jeremiah.joaquin@dlsu.edu.ph

experimental turn,[2] experimental philosophical logic seeks to employ not only the available tools and methods of traditional philosophy, but also the tools and data from the experimental cognitive sciences (especially, the psychology of reasoning) to aid philosophical inquiries into the nature of logic itself. Furthermore, some experimentally-inclined philosophers of logic envision that such a sharing of resources might lead to a more fruitful study of human cognition and reasoning (Dutilh Novaes 2012; Rips 2008; Stenning and Van Lambalgen 2008; and van Benthem 2008).

In this paper, I explore the prospects for this kind of project. In particular, I focus on two interrelated issues about the relationship between logic and the philosophy of logic, on the one hand, and the experimental results from the cognitive sciences, on the other. I cash out these issues in terms of the respective roles that logic and experiment play in the research projects in experimental philosophical logic. The first issue centers on the role of pure logic and the philosophy of logic in these projects; the second focuses on the role of these experimental results in pure logic and the philosophy of logic.

I argue for two points. First, while there are reasons to think that logic plays a crucial role in experimental philosophical logic, it is debatable whether such experimental results would impact research projects in pure logic. Though this is the case, these results could still offer some empirically backed-up data that could inform substantive debates in the philosophical logic.[3]

(Dutilh Novaes 2012), (Geurts and van Der Slik 2005), (Ghosh, Meijering, and Verbrugge 2014), and (Ripley 2011).

[2] Proponents of this *experimental turn* argue that philosophers should "proceed by conducting experimental investigations of the psychological processes underlying people's intuitions about central philosophical issues" (Knobe and Nichols 2007, 3).

[3] As was pointed out by a referee of this journal, the phrase, "philosophical logic" is ambiguous. It might refer to logic as applied to philosophy; for example, the use of possible world semantics in analyzing modality. [See (Sider 2010) for other ways that logic might be applied to philosophy.] It might also refer to philosophical issues about logic; for example, the issue about the nature of logical consequence. [See (Read 1995) for discussions on other philosophical issues in logic.] For our purposes, we follow (Restall and Russell 2012), and take philosophical logic as "the study of

The next section is scene-setting. I discuss, in broad strokes, what research projects in experimental philosophical logic are deemed to be, and how they are to be understood vis-à-vis works in pure logic and applied logic. The third section centers on the role of logic in experimental philosophical logic. I explore two possible roles for logic: a normative role, which depicts logic as a body of principles for correct reasoning; and, a descriptive role, which depicts logic as a way of modeling some target cognitive or reasoning phenomena. I show some criticisms against these depictions, and suggest that a way to address these criticisms might be by adding a prescriptive role for logic.[4]

The fourth section focuses on the issue about the role of experimental results in logic. I discuss two negative views due to L. Jonathan Cohen (1981) and Gilbert Harman (1986). I reply that though experimental results might not substantially affect work in pure logic, they could nonetheless impact work in the philosophy of logic. Finally, the final section explores two strands of work in experimental philosophical logic as case studies. I discuss first the work of Catarina Dutilh Novaes on the role of formal logic in the psychology of cognitive biases; second, I discuss the growing research on the logic and psychology of vagueness.[5] I conclude with some remarks about the importance of these works on the philosophy of logic.

logic—itself understood broadly—and its applications, pursued to philosophical ends." As such, for our purposes, we use "philosophy of logic" and "philosophical logic" interchangeably. Note, however, that this usage deviates from (Burgess 2009), (Cook 2001), (Jacquette 2007), and (Sainsbury 2001).

[4] Incidentally, the normative, descriptive, and prescriptive roles are already acknowledged in the psychology of reasoning literature [see e.g. (Bell, Raiffa, and Tversky 1988), (Evans and Over 1996), and (Stanovich 1999)]. Note, however, the nature of these roles is not uncontroversial.

[5] The literature for these two case studies is vast. Since we will just explore them for illustrative purposes, we will limit the discussion to the following works. For the first case, we focus on (Dutilh Novaes and Reck 2017) and (Dutilh Novaes 2012, 2015). For the second case, we focus on (Alxatib and Pelletier 2011), (Bonini et al. 1999), (Cobreros et al. 2012), and (Ripley 2011).

2. Pure and applied logic

Research works in logic fall under two broad categories (Ripley 2016; Priest 2005).[6] There are works in *pure* logic and there are works in *applied* logic. Works in pure logic are deeply abstract and might be classified as works in some sub-branch of pure mathematics. These works are "explorations of the properties and relations occupied by logical systems in themselves, without attending to any particular use they may or may not have" (Ripley 2016, 524). Typically, works in pure logic center on the construction and investigation of different formal systems along with their respective proof and model theories, and their respective completeness and soundness theorems. These are pursued for their own sake with no particular application in mind.

Works in applied logic, on the other hand, are more concrete, and could be seen as works that use logic for a specific purpose. Examples of such works abound. The Quine-McCluskey algorithm, for example, which is a method for determining the minimum number of Boolean functions in a given logical system, was employed in computer technology to simplify electronic circuits (Roth and Kinney 2014). The Lambek Calculus, a logic developed to analyze hierarchy of types, on the other hand, was employed in the analysis of certain grammatical structures of some natural languages (Priest 2005, sec. 12.2). Though the application of logic in these areas is of interest, we shall not be concerned with this sense of applied logic in this paper; what we shall be concerned with, rather, is arguably the canonical application of logic. Furthermore, we will look at how this application connects with one important issue in the philosophical logic, viz. whether there is one true logic.

The canonical application of logic is, arguably, in the analysis of reasoning (Priest 2005, sec. 12.4).[7] It is the science that inquires about what follows from what—i.e. what conclusion *validly* follows from a given set of

[6] I am aware that there are other ways of classifying works in logic. But this broad classification of pure and applied logics would suffice for my purposes here.

[7] As was pointed out by this journal's referee, this point is controversial. As we will see later in sec. 3, some theorists suggest a distinction between logic as concerned

premises.[8] And like many other types of scientific inquiry, logic is couched in a highly mathematical, formal language. This formalization, in turn, provides an account of validity.

But there is a complication here. As it turns out, there are many logics. There are many formal systems of logic, each of which offers a different account of validity. For example, classical logic provides a notion of validity in terms of the impossibility of the conclusion to be false while the premises are true. Intuitionistic logic, on the other hand, gives an account of validity in terms of constructive proofs from a given set of premises to a conclusion. Some paraconsistent logics offer still another different conception of validity in terms of some type of relevance relation between premises and conclusions (Beall and Restall 2006). And there are more. In fact, there is a host of other pure logics, each of which offers a different formal account of validity. Some of these formal accounts cohere with one another; others conflict with each other. This plurality of logics leads to a central issue in discussions in the philosophy of logic.

One central issue in the philosophy of logic precisely amounts to the issue of whether there is only one true logic, or whether there are many, equally true logics (Beall and Restall 2006; Russell 2008). By "logic" we mean a formal theory of consequence (i.e. deductive validity). By "true logic" we mean the *correct* account of logical consequence. Monists, like Priest (2005), argue that there can only be one true account of logical consequence. Pluralists, like Beall and Restall (2006), on the other hand, argue that there are many, equally true logics.[9]

The issue between pluralists and monists about logical consequence is a broad and delicate philosophical issue. It is broad in that the debate involves a lot of aspects of logic. It is not only concerned about the nature of logic itself, but also of the nature of logical constants, logical truths, and logical consequence. It is delicate in that what is at stake is the core of

about deductive validity as opposed to reasoning *per se*, see e.g. (Harman 1986); for a reply, see (Field 2009) and (Sainsbury 2002).

[8] Throughout this paper, we will be concerned with the notion of *deductive* validity and not *inductive* validity.

[9] Perhaps, there is a third, ultra negative view, logical nihilism, which tells us that there's no such thing as a logic (Russell 2017).

reasoning itself. If it turns out that monism is right, then there is only one correct way of reasoning. If it turns out the pluralism is right, then there is more than one way of reasoning correctly.

There is no fast way of adjudicating between pluralism and monism. But the issue could be made tractable by translating the issue into the question of the adequacy of a logical theory to account for its data. Given the canonical application of logic, a logical theory ought to provide an account of correct reasoning. This account, however, need not be a general model of reasoning; it might just be about a particular piece of reasoning. To check whether some logical theory is correct, we need to check whether it adequately accounts for the data it purports to explain.

Consider, for example, how some logical theory explains how we reason about vagueness. Take any logical theory that models vagueness, for example, Lukasiewicz's three-valued logic, L3, or Priest's Logic of Paradox, LP. Check whether the proffered logical theory adequately explains not only the target phenomenon, but also the sorites paradox, which results from it.[10]

This strategy connects up nicely with the pluralism/monism debate. If there are two or more logical theories adequately explain how we reason about vagueness, then, perhaps, all things being equal, taking a pluralist view about the logic of vagueness is the right way to go. On the other hand, if only one theory is adequate, then, perhaps, monism is the right view. If none of the current theories accounts for vagueness, however, then we should perhaps consider other logics to account for it.

Determining the adequacy of a given logical theory, however, is not a straightforward matter, and has been a subject of much debate. Some theorists have suggested that, at the very minimum, logical theories should provide an accurate formalization of the target phenomenon in that the *relevant* aspects of the target phenomenon are formalized appropriately. Others put premium to the *informativeness* of the proffered theory. A given logical theory should be able to provide a non-circular, non-trivial explanation of the target phenomenon. But whatever the criterion for adequacy might be, a logical theory must be able to explain its data. That is, it must be able to explain the piece of reasoning it aims to account for. In a way,

[10] For a useful introduction to the various logics of vagueness, see (Williamson 1995) and (Smith 2008).

one might argue that the criteria for adequacy of a given logic should be based on the same set of criteria used in the sciences, like physics, chemistry, and biology, viz., simplicity, fruitfulness, explanatory power, etc.[11]

Treating different logics as scientific theories naturally leads us to the question about the importance of experimental data to logic. The hard sciences give importance to experimental data as they are employed as a test to either confirm or disconfirm a proposed physical theory. For example, Einsteinian relativity theory was confirmed by some set of experimental data due to Arthur Eddington; much in the same way that the phlogiston theory was disconfirmed by some other set of experimental data concerning heat and its effect to molecular movement. But even if we grant that experimental data are important in these fields, is this true of theories of logic as well? Are experimental data important to logic? Experimentally-minded philosophers and logicians think that this is so.

Like many philosophers who advocate an experimental turn in philosophy, experimentally-minded philosophers and logicians think that experimental data are important to logic, especially, to the philosophy of logic, since they could serve as checks on philosophical (logical) intuitions; and, as a test of a philosophical (logical) theory's empirical claims (Chalmers 2009, 2007).[12] For example, as experimental moral philosophers test whether our actual moral intuitions are consistent with the claims of some moral theory experimental logicians test whether our actual judgments and reasoning processes are best explained by some logical theory. This implies, then, that experimental data might serve as a check for the empirical adequacy of a given logical theory. Thus, given that logical theories aim to account for certain forms of reasoning, experiments could be set up to test whether such theories actually provide accurate descriptions of them.[13]

[11] See, for example, the discussion of Alfred Tarski's and Rudolf Carnap's criteria for formal adequacy in (Dutilh Novaes and Reck 2017); for a discussion of Carnap's views of definitional adequacy, see (Chalmers 2012, chap. 1); for a discussion of the 'science-like' mechanisms for theory-selection in logic, see (Priest 2014; 2005, chap. 8).

[12] Permission to cite these two references was granted by David Chalmers.

[13] Works cited in footnote 1 exemplify this.

To illustrate, let us take classical logic as our target formal system, and check whether its inference rules hold in ordinary reasoning. One controversial inference rule in classical logic is explosion: from a contradiction, anything follows. Experimentally-minded logicians might want to test whether this principle holds in *actual* reasoning. They may do this by conducting experiments to test a subject's inferential responses to contradictory information.

For example, subjects might be tasked to deduce certain conclusions from a given set of facts which include contradictory information. Suppose the set includes the following: "Amanda is a grade 6 student;" "Amanda is 11 years old;" "If Amanda is a grade 6 student, then she is smart;" and "Amanda is not 11 years old." Experimenters could then ask their subjects whether the following conclusions, "Amanda is smart" and "Manila is in the Philippines," could be validly inferred from the set. If the subjects answer "Yes" to both, then we could say that they are abiding by the principles of classical logic.[14] If they only answer "Yes" to the former, then we could say that they are employing a different, non-explosive logic in making their inferences; this, perhaps, is a kind of paraconsistent, relevant logic.[15]

As such, in this picture, experimental data is employed to test and validate whether actual reasoners employ some non-classical logical device in their inferences. Furthermore, the logical device itself would serve as an explanation of why reasoners make inferences the way they do.

Experimental philosophical logic works by exploring two important features: (1) that a formal model can account for a target phenomenon; (2) that the target phenomenon being modeled is itself amenable to experimental exploration. The first feature implies that the target phenomenon,

[14] This is so since, via modus ponens, from "If Amanda is a grade 6 student, then Amanda is smart" and "Amanda is a grade 6 student," we could validly infer that "Amanda is smart." And via explosion, the conclusion, "Manila is in the Philippines" could be validly inferred from the contradictory premises: "Amanda is 11 years old" and "Amanda is not 11 years old."

[15] This is so since, though explosion is invalid in a paraconsistent logic, it still permits certain valid inferences so long as they do not include contradictory premises as part of the inference.

be it linguistic behavior, social communicative behaviors, etc., can be modeled by a given formal system. The second feature, on the other hand, implies that there is an experimental way to validate (or verify) whether the target phenomenon is indeed modeled by the formal theory (Ripley 2016). The combination of these two aspects will tell us whether a given logical theory is empirically adequate; it would also tell us whether the theory is a good scientific explanation of a piece of reasoning.

One may notice, however, that there are key assumptions to this experimental approach to philosophical logic. On the one hand, it treats logical theories not just as normative theories of how reasoners *ought* to think, but as descriptive theories of how people *actually* think. On the other hand, it treats experimental data as genuine tests for the adequacy of logical theories. One might question these assumptions; and thus might put into question the whole enterprise of an experimental approach to philosophical logic. To question the first assumption is to question the role of logic in experimental philosophical logic; to question the second is to question the role of experimental data in the enterprise. We will look at these issues in the following two sections.

3. Three roles for logic: descriptive, normative, and prescriptive

One issue that can be raised about projects in experimental philosophical logic has something to do with the role of logic in these projects. What role, if any, does logic play in experimental philosophical logic? There are two prominent views about the role of logic found in the literature: a normative role and a descriptive role (Stich 1990, 13–16).

On the normative view, a logical theory is taken as a body of reasoning principles that serves as the *norms* for correct reasoning. Logical principles function as a kind of permission inference-tickets; i.e. as rules that tell us that from a given premise, such and such a conclusion follows. Alternatively, they can be seen as rules of inference that specify the conditions for validity; i.e. as structural rules that tell us that if some natural language argument is an instance of such and such an argument structure, then it is valid. For example, classical logic has a disjunction-introduction rule that tells us that from any proposition, P, one could infer the disjunction, P or Q. Thus, from the premise “Hazel is loving” we could validly infer that “Either Hazel is loving or Hazel is sweet.” Alternatively, since the argument, “Hazel is loving; therefore, either Hazel is loving or Hazel is sweet” is an instance of the disjunction-introduction rule, we could conclude that such an argument is valid.

On the descriptive view, a logical theory is taken as a model (i.e. a formal description) of a given reasoning phenomena. According to this view, logic is a kind of zoological study whose primary task is to document, formalize, and categorize various types of reasoning in an appropriate logical system. For example, arguments, like “If Candice is 6 years old, then Candice is no longer a baby; Candice is 6 years old; thus: Candice is no longer a baby” is piece of valid reasoning. Arguments with same structure preserve truth from premise to conclusion. The task then of a logical theory is formalize this structure into something known as *modus ponens*, “If P, then Q; P; therefore, Q.” This formalized structure would, then, be treated as a cognitive artifact[16] which would be catalogued and indexed to some logical system.

Each of these roles, however, seems to imply certain worries. Treating logic as a system that provides norms for correct reasoning seems to overestimate the capacity of actual (human) reasoners. That is, taken by its normative role, experimental philosophers of logic might evaluate reasoners in terms of some standard logic—usually, in terms of classical logic; since experimental data have shown that actual reasoners do not always follow the norms of this standard logic, these philosophers might make the conclusion that human beings are generally just poor reasoners, or are always prone to systematic errors (Evans and Over 1996, 4). This is worrisome because it might be the case that some people employ a different kind of logic that does not abide by the principles of classical logic.

Treating logic as descriptive models of human reasoning, on the other hand, seems to deny the normative force of logical principles. Focusing mainly on the descriptive role of logic, experimental philosophers of logic

[16] A cognitive artifact is defined as a physical extension or manifestation of our cognitive abilities (Dutilh Novaes 2012).

might be overwhelmed by the different types of reasoning deployed by ordinary reasoners such that they will just take the data as they come, and catalogue them in neat logical systems. But this would mean that these “logics” or forms of reasoning could not be evaluated as good or bad since each of them obeys some kind of “logic”. Thus, a descriptive view of logic might lead to a kind of logical relativity where each form of reasoning would be as good as any other.

These worries are important methodological issues about the role of logic in experimental philosophical logic. And it is important to address them if we are to see the fruits of projects in this area. Furthermore, addressing these worries might lead to a *reflective equilibrium* where the descriptive and normative roles are balanced out (Dutilh Novaes 2012, 79).

Some initial steps to address these worries, however, could already be seen in the psychology of reasoning literature. For example, some theorists have suggested distinguishing the normative role of logic (and other normative systems of reasoning, e.g. probability and decision theory) from its prescriptive role. In this prescriptive view, logical theories are seen as reasoning prescriptions for actual, real-life reasoning. The rules of inference embedded in a logical theory are taken not as norms for correct reasoning that cuts across various reasoning contexts; rather, they are taken as guidelines of how to reason in particular reasoning contexts.[17] For example, instead of thinking of Priest’s LP as a universal norm for reasoning, we might take it as a prescription that when we reason about the liar sentence, “This sentence is false,” we ought to take it as both true and false. Likewise, if we are dealing with vague sentences, like sentences about future events, we might take Łukasiewicz’s L3 as a prescription that we ought to judge them as neither true nor false. Taken this way, each logical system would prescribe correct forms of reasoning in particular contexts. As characterized thus far, the prescriptive role of logic would be a kind of normative logical pluralism, while the normative role would be a kind of normative logical monism.[18]

[17] This point, I think, is shared by Dutilh Novaes (2012); see also (Stanovich 1999) and (Bell, Raiffa and Tversky 1988).

[18] This point was raised by a referee of this journal.

But what is going for the prescriptive role of logic is that it has two theoretical merits. First, it recognizes that human reasoners are not ideal rational reasoners. They need to be told how to reason in particular circumstances. In this way, the prescriptive view has sidestepped the issue raised against the normative view. Second, it also recognizes the normative force of logical principles; since prescriptions are context-relative "oughts," they already imply some kind of normative force. In this way, the prescriptive view has also addressed the worry raised against the descriptive view.

One further merit of disentangling the prescriptive role from the normative role of logic is that it gives us a new way of viewing of the interplay between the three roles of logic. Furthermore, we will have an appreciation of the different kinds of projects for experimental philosophers of logic; which further implies different tests of adequacy for these projects.

In assessing logic's normative role, for example, experimental philosophers of logic (and logicians in general) are more concerned with a logical theory's internal, theoretical merits rather than its applications. As such, logicians are more concerned about the *theoretical* adequacy of the theory, where simplicity, non-*ad hocness*, etc. would be factors. In assessing logic's descriptive role, on the other hand, experimental philosophers of logic are concerned about the *empirical* adequacy of a logical theory. That is, their main concern would be how well the theory accounts for the data observed in reasoning experiments.[19] Finally, in assessing logic's prescriptive role, experimental philosophers of logic are more concerned about the utility or pragmatic value of these reasoning prescriptions. Their main concern would revolve around issues about the applicability of logical devices in specific reasoning circumstances. Thus, in this tripartite view of logic, different roles imply different projects, and different tests for adequacy.

Though adding a prescriptive role to logic might have its merits, it is still susceptible to some other worries. For example, given that logical theories are reasoning prescriptions, there is a worry of how to decide what logical theory should be prescribed in a given reasoning context. Furthermore, there is a worry about who should make such decisions. There is

[19] This is, perhaps, one way of appreciating the current literature in the psychology and philosophy of vagueness.

another more fundamental worry concerning the *ad-hocness* of the distinction between the normative and prescriptive roles. If this distinction does not hold, then all the merits we have seen would be put into question. These are further challenges that experimentally-inclined philosophers of logic need to meet.

4. Experimental data and philosophical logic

Another issue for an experimental approach to philosophical logic is about the role and importance of experimental data in logic itself. Two negative views could be cited here. One comes from L. J. Cohen (1981); the other from Gilbert Harman (1986).[20] Both imply the independence of experimental data and logic from one another.

Cohen argues that there seems to be a gap between experimental data and logic. Though experimental data might indicate that many reasoners perform poorly in some reasoning tasks, it does not follow that they are incompetent in terms of the implied norms of a given logical theory. The reasoners could just be employing some other system of logic. For example, data from the Wason card-selection experiment have shown that human reasoners are poor at judging conditional statements (Ripley 2016; Joaquin and Agregado 2018). But from these data, Cohen argues, we could not make any evaluation of the adequacy of the implied normative, classical logical system assumed in the experiments since such a logical framework is just assumed by the experimenters. Nor could we make any evaluation of the human subjects' reasoning competencies since the human subjects might just be employing a different rule for conditional reasoning.

In a slightly different angle, Harman seems to echo the same point. We could take him as arguing that since reasoning and logic are about two different things, and are governed by two different sets of considerations, it

[20] These two objections are already well-documented in the literature. For example, Evans and Over (1996) discuss Cohen's views extensively; Harman's view of logic and its relation to reasoning, and the commentaries of other philosophers are detailed in (Dutilh Novaes 2015). The discussion here will just highlight the salient points regarding the role of experimental data in logic.

follows that experimental results about reasoning must be evaluated independent of the correctness of a given logic.

For Harman, logic is concerned with implications, with consequence relations, with what follows from what. For example, in logic, we are concerned with whether some statement, Q, follows from (or is implied by, or is a consequence of) the statements: If P, then Q; and, P. And that's the end of that. Reasoning, on the other hand, is concerned with the reasonability of belief revision, with what he calls, "reasonable change in view." For example, in everyday reasoning, you might *infer* (in the sense of that you cognitively suppose) that there is milk in the fridge from the fact that you bought one yesterday. However, finding out that there is no milk in the fridge might make you revise some of your starting beliefs. Perhaps, you might now suppose that you forgot to buy milk yesterday, or that you have just misplaced it, or that some extraterrestrial alien took it. Of the three choices here, we might judge the last as the most unreasonable, while the other two as reasonable. Of course, the second would be more reasonable if you find the receipt that proves that you bought milk yesterday. In judging the reasonability of these options, however, logic plays little to no role at all. What is doing the work here has something to do with epistemic considerations about the reliability of our evidence (in the case of the receipt) and some sort of background knowledge (in the case of extra-terrestrial aliens).

For Harman (1986, 11–12), the assessment of the reasonability of some belief, Q, does not stem from the principles of logic. Following the principles of logic, Q might be logically deduced from "If P, then Q" and "P." But even if one believes the starting information and the validity of the inference, one might still not come believe Q. Thus, the assessment of the correctness of a given logical theory seems to be independent of the reasonability of belief revision since the latter is governed by a set of epistemic considerations, which do not really govern the former.

Both Cohen's and Harman's views seem to imply that our evaluation of the actual reasoning performance of human subjects (found in reasoning experiments) must be independent from our evaluation of principles of some given logic. For Cohen, this is so since human subjects might be using a different logic in performing their tasks. For Harman, this is so since reasoning and logic are about two different things. Having said this, it is now useful to divide the issue about the role of experimental data in logic into two sub-issues: first, the issue about their role in *pure logic*; second, their role in *philosophical logic*, and see whether both of their views work for these sub-issues.

At the onset, it would seem that experimental data do not really play any role in pure logic. But this is so independent of the views of Cohen or Harman. Recall that we have characterized works in pure logic as highly abstract, and said that they are more concerned with the formal features of logical systems rather than their applications. Viewed this way, these works are evaluated in terms of their theoretical adequacy much like we assess the normative role of logic discussed above. Furthermore, logicians evaluate a logical theory's elegance and simplicity much like mathematicians judge the elegance and simplicity of mathematical proofs and theories.

Experimental results, on other hand, seem to have a role in philosophical logic given that the canonical application of logic is in the analysis of reasoning. As such, the views of Cohen and Harman might weigh in. To address Cohen's view, an experimental philosopher of logic might reply as follows. Suppose that Cohen is right that human subjects do employ different types of logics in performing reasoning tasks. Then all the more an experimental approach to philosophical logic should be undertaken in order to identify these sorts of logics.

To address Harman's view, on the other hand, our experimental philosopher of logic might reply as follows. Harman is surely wrong that logic is just concerned with (logical) implications since there are other logics, like some non-monotonic logics, which are concerned with belief revision as well [see, e.g., (Dutilh Novaes and Veluwenkamp 2017)]. At least in these types of logics, experimental data are needed as validation.[21]

5. Two case studies

As a conclusion, in this final section, we will explore two strands of current work in experimental philosophical logic to illustrate the fruitfulness

[21] See (Dutilh Novaes 2015, 590–91) for further clarifications about this matter. See also (Field 2009) for a different reply to Harman.

of this kind of approach. First, we explore the work done by Dutilh Novaes about the potential prescriptive role that logic plays in the psychology of cognitive biases; second we explore work done in the fast-growing literature on the psychology of vagueness, which takes experimental data to evidence philosophical/logical theories of vagueness.[22]

Dutilh Novaes (2012) investigates the role that formal languages play in the psychology of reasoning, esp. in the psychology of cognitive biases. Cognitive biases are reasoning mistakes that people often commit; they are "systematic errors" which are brought about by our limited cognitive capacities. One particular cognitive bias that she focuses on is belief bias, i.e. the tendency to take logically invalid arguments with believable conclusions as valid.

Belief bias is a well-documented phenomenon in the psychology of reasoning literature. A series of experiments on syllogistic reasoning competence has shown that many people would endorse the validity of a logically invalid syllogism if its conclusion is believable and the concepts used therein are comprehensible. For example, only 32% of the test subjects endorsed the logically correct indictment (i.e., the argument is invalid) when confronted with this syllogism:

> All living things need water. Roses need water. Therefore, roses are living things.

On the other hand, confronted with an argument with the same invalid syllogistic form, but phrased in unknown concepts, viz.:

> All animals of the hudon class are ferocious. Wampets are ferocious. Therefore, wampets are animals of the hudon class.

78% of the same test subjects gave the logically correct answer (Dutilh Novaes 2012, 94).

Dutilh Novaes conjectured that what is doing the work in the first case is our tendency to have an automatic response to known data; while in the second case it is our tendency to slow down our thinking when it comes to unknown data. In the former case, we are susceptible to belief bias. In the

[22] For references, see note 5.

latter case, the principles of logic factor in our reasoning process. Furthermore, she suggests that in order to improve our reasoning skills, we ought to take a formal language (like, the language of syllogistic logic) as a tool to counterbalance our default reasoning processes, such as belief bias. In this way, we could interpret Dutilh Novaes as providing a prescriptive role for logic.[23]

If in Dutilh Novaes' work, logic plays a potential prescriptive role, in the works in the logic and psychology of vagueness, logic plays a more descriptive role. Recall that seeing logic in its descriptive role implies seeing it as a kind of study whose primary task is to categorize various types of reasoning in an appropriate formal system. This can be seen in the way the literature on vagueness has developed in the last few years.

Research on vagueness has led to the creation (or discovery) of philosophically interesting logical systems, which aim to explain the phenomenon. But since the 1990's there has been a steady growth of philosophical studies which employed experimental data to evidence logically-couched philosophical theories.

Bonini and his colleagues (Bonini et al. 1999), for example, present experimental results about the use of vague expressions by native Italian speakers, and take the results to count in favor of a "vagueness-as-ignorance" view—an *epistemicist* theory of vagueness, which tells us that vagueness (i.e., expressions which seem to lack sharp boundaries or have truth-value gaps) only occurs because we lack the knowledge of the actual boundaries of concepts we employ. In one of their experiments, subjects were tasked to fill-in a questionnaire which tests their judgments about tallness. They were asked questions (in Italian) amounting to:

[23] A referee of this journal pointed out that Dutilh Novaes' view seems to imply that a given logic is not an instruction but a tool used to counterbalance our default reasoning tendencies. As such, logic does not really prescribe a set of rules for correct reasoning. I reply, however, that as a tool, logic does come with a prescribed set rules for correct reasoning that may be effectively used in certain reasoning contexts. As Dutilh Novaes (2012, 3) tells us, "the historical development of formal languages can be viewed as a process of cultural evolution through which humans looked for tools that would allow them to perform certain tasks and solve certain problems more efficiently [...]."

A man is tall if his height is greater than or equal to ___ centimeters.

A man is not tall if his height is lesser than or equal to ___ centimeters.

A man is at least of average height among 30-year-old Italians if his height is greater than or equal to ___ centimeters.

Surprisingly, the results have shown that "judgments of the lower threshold it takes to be tall were significantly higher than judgments of the higher threshold it takes to be not tall" indicating truth-value gaps, even though the two latter questions seem to imply the existence of actual boundaries between tall and not tall average 30-year-old Italians.

Several years later, other experimental studies have shown a different and a more paraconsistent-friendly result. For example, Ripley (2011) tested how subjects appreciate and evaluate the vague relational predicate, "near," and found that when it comes to borderline cases, subjects tend to tolerate contradictions of the form, x is both near and not near to y. The experiment goes this way. Subjects were shown seven pairs of figures (A to G) each consisting of a square and a circle at decreasing distances from each other. The extreme cases, A and G, are clear-cut cases. Case A was a clear case where the square and the circle are far apart, while Case G shows these figures as clearly close to each other. The less extreme cases, B and F, showcase a little decrease (in the case of B) and a little increase (in the case of F) of distances of the figures. And the borderline cases, C to E, are the target cases. Subjects were, then, asked whether they agree that the contradictory sentence "the circle is near the square and it isn't near the square," along with its linguistic variants, is true. What was found is that a significantly greater proportion of subjects fully agree with the contradictory sentence as they approach the borderline cases than for the extreme cases. This evidences a kind of paraconsistent logic, which accepts true contradictions.[24]

Further work in the logic and psychology of vagueness has already been pursued, and has led to the developments of certain logics. For example, Cobreros et al. (2012) have developed a strict-tolerant logic based on three

[24] Alxatib and Pelletier (2011) also present the same result, but from a different experimental set-up.

notions of truth: classical truth, strict truth, and tolerant truth. This logic aims not only to account for vagueness, but also to account for the proffered experimental data. The logic defines up strict truth (akin to Kleene's K3 logic) and tolerant truth (akin to Priest's LP), and shows that experimental results could be explained in terms of these two dual notions.

The field of inquiry into the logical and psychological aspects of vagueness is quite open; and it looks out for new and exciting ways of building-up the relationship between logic and experiment. But this is not only true of research in vagueness. The interaction between the two might yield more interesting results in areas where reasoning and logic coincide. As one experimental philosopher of logic notes:

> We should expect experiment and logic to fruitfully interact whenever a field of inquiry involves rigging up a logical system to capture some experimentally explorable phenomenon; in these cases, logical approaches will help us decide which aspects of the phenomenon to experimentally explore, and experimental approaches will help us choose which logics best capture the phenomenon. (Ripley 2016, 533)

And this is something experimentally-minded philosophers of logic are keen to do and achieve.

Acknowledgements

A version of this paper was presented at the Australasian Association for Logic 2015 Conference. I would like to thank the participants and organizers, especially Nick J. J. Smith and Simon Varey, for their useful feedback. Special thanks to Hazel Biana, Robert James Boyles, Ben Blumson, David Chalmers, Mark Anthony Dacela, Catarina Dutilh Novaes, James Franklin, Adrianne John Galang, Napoleon Mabaquiao, Graham Priest, and David Ripley for the discussions that helped shape the key themes of this paper.

References

Alxatib, Sam, and Francis Jeffry Pelletier. 2011. "The Psychology of Vagueness: Borderline Cases and Contradictions." *Mind & Language* 26 (3): 287–326. https://doi.org/10.1111/j.1468-0017.2011.01419.x

Beall, Jeffrey, and Greg Restall. 2006. *Logical Pluralism*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199288403.001.0001

Bell, David E., Howard Raiffa, and Amos Tversky, eds. 1988. *Decision Making: Descriptive, Normative, and Prescriptive Interactions*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511598951

Bonini, Nicolao, Daniel Osherson, Riccardo Viale, and Timothy Williamson. 1999. "On the Psychology of Vague Predicates." *Mind & Language* 14 (4): 377–93. https://doi.org/10.1111/1468-0017.00117

Braüner, Torben. 2014. "Hybrid-Logical Reasoning in False-Belief Tasks." In *Logic and Interactive Rationality, Volume II: Yearbook 2012*, edited by Zoé Christoff, Paolo Galeazzi, Nina Gierasimczuk, Alexandru Marcoci, and Sonja Smets, 79–103. University of Amsterdam.

Burgess, John P. 2009. *Philosophical Logic*. Princeton: Princeton University Press.

Cobreros, Pablo, Paul Egré, David Ripley, and Robert van Rooij. 2012. "Tolerant, Classical, Strict." *Journal of Philosophical Logic* 41 (2): 347–85. https://doi.org/10.1007/s10992-010-9165-z

Cohen, L. Jonathan 1981. "Can Human Irrationality be Experimentally Demonstrated?" *Behavioral and Brain Sciences* 4 (3): 317–31. https://doi.org/10.1017/S0140525X00009092

Cook, Roy T. 2009. *Dictionary of Philosophical Logic*. Edinburgh: Edinburgh University Press.

Chalmers, David J. 2007. "X-Phi Meets A-Phi." Talk presented at Experimental Philosophy Meets Conceptual Analysis, slides, URL: http://consc.net/papers/xphi.ppt

Chalmers, David J. 2009. "What can Experimental Philosophy Do?" Talk presented at NEH Institute on Experimental Philosophy, slides, URL: http://consc.net/papers/xphil.ppt

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

Dutilh Novaes, Catarina. 2012a. *Formal Languages in Logic: A Philosophical and Cognitive Analysis*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139108010

Dutilh Novaes, Catarina. 2012b. "Towards a Practice-Based Philosophy of Logic: Formal Languages as a Case Study." *Philosophia Scientiae* 16 (1): 71–102. https://doi.org/10.4000/philosophiascientiae.719

Dutilh Novaes, Catarina. 2015. "A Dialogical, Multi-Agent Account of the Normativity of Logic." *Dialectica* 69 (4): 587–609. https://doi.org/10.1111/1746-8361.12118

Dutilh Novaes, Catarina, and Erich Reck. 2017. "Carnapian Explication, Formalisms as Cognitive Tools, and the Paradox of Adequate Formalization." *Synthese* 194 (1): 195–215. https://doi.org/10.1007/s11229-015-0816-z

Dutilh Novaes, Catarina, and Herman Veluwenkamp. 2017. “Reasoning Biases, Non-Monotonic Logics and Belief Revision.” *Theoria* 83 (1): 29–52. https://doi.org/10.1111/theo.12108

Evans, Jonathan S.B., and David E. Over. 2013. *Rationality and Reasoning.* East Sussex: Psychology Press.

Field, Hartry. 2009. “What is the Normative Role of Logic?” *Proceedings of the Aristotelian Society* 83 (1): 251–68. https://doi.org/10.1111/j.1467-8349.2009.00181.x

Geurts, Bart, and Frans van der Slik. 2005. “Monotonicity and Processing Load.” *Journal of Semantics* 22 (1): 97–117. https://doi.org/10.1093/jos/ffh018

Ghosh, Sujata, Ben Meijering, and Rineke Verbrugge. 2014. “Strategic Reasoning: Building Cognitive Models from Logical Formulas.” *Journal of Logic, Language and Information* 23 (1): 1–29. https://doi.org/10.1007/s10849-014-9196-x

Harman, Gilbert. 1986. *Change in View.* Cambridge, Mass.: MIT Press.

Jacquette, Dale, ed. 2006. *A Companion to Philosophical Logic.* London: Blackwell Publishing. https://doi.org/10.1002/9780470996751

Joaquin, Jeremiah Joven, and Jose Emmanuel Agregado. 2018. “Grounding Logic: A Reply to Shenefelt and White.” *Think: Philosophy for Everyone* 17 (49): 13–16. https://doi.org/10.1017/S1477175618000052

Knobe, Joshua, and Shaun Nichols. 2007. “An Experimental Philosophy Manifesto.” In *Experimental Philosophy*, edited by Joshua Knobe and Shaun Nichols, 3–14. Oxford: Oxford University Press.

Priest, Graham. 2005. *Doubt Truth to be a Liar.* Oxford: Oxford University Press. https://doi.org/10.1093/0199263280.001.0001

Priest, Graham. 2014. “Revising Logic.” In *The Metaphysics of Logic*, edited by Penelope Rush, 211–23. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139626279.016

Read, Stephen. 1995. *Thinking about Logic.* Oxford: Oxford University Press.

Restall, Greg, and Gillian Russell. 2012. *New Waves in Philosophical Logic.* New York: Palgrave Macmillan. https://doi.org/10.1057/9781137003720

Ripley, David. 2011. “Contradictions at the Borders.” In *Vagueness in Communication*, edited by Rick Nouwen, Robert van Rooij, Uli Sauerland, and Hans-Christian Schmitz, 169–88. New York: Springer. https://doi.org/10.1007/978-3-642-18446-8_10

Ripley, David. 2016. “Experimental Philosophical Logic.” In *Blackwell Companion to Experimental Philosophy*, edited by Wesley Buckwalter and Justin Sytsma, 523–34. London: Wiley-Blackwell. https://doi.org/10.1002/9781118661666.ch36

Rips, Lance J. 2008. “Logical Approaches to Human Deductive Reasoning.” In *Reasoning*, edited by Jonathan E. Adler and Lance J. Rips, 187–205. Cambridge: Cambridge University Press.

Roth, Charles H., and Larry L. Kinney. 2014. *Fundamentals of Logic Design*, 7th edition. Stamford, CT: Cengage Learning.

Russell, Gillian. 2008. "One True Logic?" *Journal of Philosophical Logic* 37 (6): 593–611. https://doi.org/10.1007/s10992-008-9082-6

Russell, Gillian. 2017. "An Introduction to Logical Nihilism." In *Logic, Methodology and Philosophy of Science – Proceedings of the 15th International Congress*, edited by Hannes Leitgeb, Ilkka Niiniluoto, Päivi Seppälä, and Elliot Sober, 120–31. London: College Publications.

Sainsbury, Robert Mark. 2001. *Logical Forms: An Introduction to Philosophical Logic*, 2nd edition. London: Blackwell.

Sainsbury, Robert Mark. 2002. "What Logic Should We Think with?" *Royal Institute of Philosophy Supplements* 51: 1–17. https://doi.org/10.1017/S1358246100008055

Sider, Theodore. 2010. *Logic for Philosophy*. Oxford: Oxford University Press.

Smith, Nicholas J.J. 2008. *Vagueness and Degrees of Truth*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199233007.001.0001

Stanovich, Keith. 1999. *Who is Rational? Studies of Individual Differences in Reasoning*. New Jersey: Lawrence Erlbaum Associates, Inc.

Stenning, Keith, and Michiel van Lambalgen. 2008. *Human Reasoning and Cognitive Science*. Cambridge, MA: MIT Press.

Stich, Stephen. 1990. *The Fragmentation of Reason*. Cambridge, MA: MIT Press.

Van Benthem, Johan. 2008. "Logic and Reasoning: Do the Facts Matter?" *Studia Logica* 88 (1): 67–84. https://doi.org/10.1007/s11225-008-9101-1

Williamson, Timothy. 1995. *Vagueness*. London: Routledge.

Organon F 26 (2) 2019: 287–297
https://doi.org/10.31577/orgf.2019.26206

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

RESEARCH ARTICLE

Casting a Shadow on Lewis's Theory of Causation

Erdinç Sayan*

Received: 2 October 2018/ Accepted: 20 November 2018

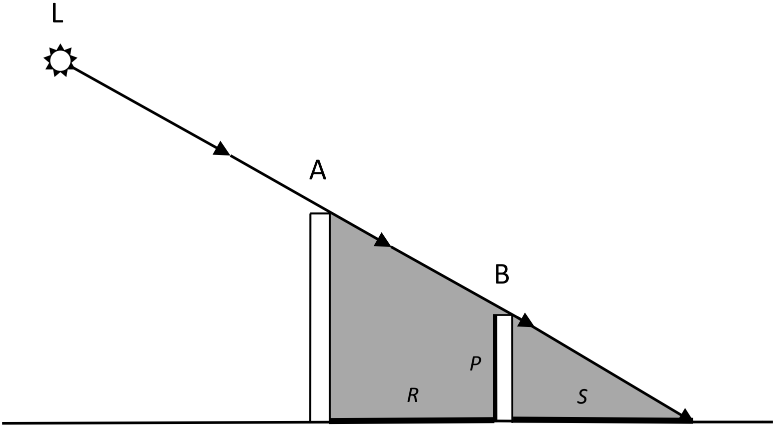
Abstract: First I present a puzzle involving two opaque objects and a shadow cast on the ground. After I offer a solution to this puzzle by identifying which of the objects is causally responsible for the shadow, I argue that this case poses a counterexample to David Lewis's latest counterfactual account of causation, known as his influence theory. Along the way, I discuss preemption, overdetermination, absence causation, and trumping preemption.

Keywords: Absence causation; counterfactual theory of causation; influence theory of causation; overdetermination; preemptive prevention; trumping pre-emption.

1.

Here is a puzzle: In the below cross-sectional diagram, L is a very distant light source (like the sun), A and B are two opaque rectangular objects with equal widths but different heights. (We can take the thicknesses of A and B as negligible.) The light ray coming from L grazes the upper right edges of A and B. If only A had been present, it would have cast the shadow $R+S$ on the ground; and if only B had been present, it would have cast the

* Middle East Technical University
✐ Philosophy Department, Middle East Technical University, Dumlupinar Bulvari, 1, 06800 Ankara, Turkey
✉ esayan@metu.edu.tr



shadow *S*. In the situation above, *P* is the shadow A casts on B (which coincides with the area of B facing A).

Question: Clearly, the shadows *R* and *P* are caused by A. But which one of A and B is causally responsible for the shadow *S*?

(1) B cannot be causing *S*, because B is not receiving any light, as A completely blocks all the light from reaching B. And an object which does not have any light impinging on it cannot cast any shadow.

(2) A cannot be causing *S*, because A's casting *S* is prevented by A's casting *P* on B. An object can cast only one shadow in the presence of one light source. In this case, A's full shadow due to L is *R*+*P*; so we cannot claim that A casts *S in addition to* casting *P*.

(3) Since neither A nor B is causally responsible for *S*, we cannot say A and B overdetermine *S*. For in overdetermination situations, there are two or more causes each of which produces the very same effect independently of the others. Nor can we say that there is preemption here—that one of A and B is preempting the other and is itself causing *S*—for neither is a cause of *S*.

Then *what* is causing *S*?

2.

My answer to this puzzle is that it is A, despite the considerations in (2). Clearly, A is what is causing the dark region (umbra) to the right of A, by blocking the light coming from L. (B has no share in bringing about that dark region, as B does not receive any light.) The presence of the ground (represented by the long horizontal line in the figure) that intersects with the dark region leads to the formation of the shadow *R*+*S* on the ground. Hence, it is A that is causally responsible for *S*. Thus we need to give up the principle we mentioned in (2), that an object can cast only one shadow in the presence of one light source.

B is a "back-up cause" of *S*: If A had not been present, B would have cast *S*. It follows that we do have a case of preemption here, after all: A preempts B from causing *S*. B cannot be said to be an overdetermining cause of *S* together with A. For to claim that A and B are overdetermining causes of *S* would be to imply that *both* A and B can be credited for causing *S* independently of the other. But B can be given no such credit, as A blocks all the light from reaching B. The causal pathway from B to *S*, which would have existed had A not been present there, is thwarted by the presence of A.

3.

I asserted above that the presence of A is the cause of the formation of *S* and that the presence of B is merely a potential cause of it which is preempted by A. In making that assertion I assumed the following description of the effect-event:

e_{fs}: formation of the shadow *S*.

But one might choose to describe the effect-event as follows, instead:

e_{pl}: prevention of light from reaching the surface *S*.

With the second description e_{pl}, the situation in the diagram becomes a case of "redundant prevention" or "preemptive prevention": B's prevention of light from entering *S*'s region was preempted, or was redundant, because of the presence of A. Let us now ask if our causal judgments above will be different if we view the situation as a case of redundant or preemptive prevention.

Let us take a closer look at the notion of redundant prevention using the following example of a redundant prevention Michael McDermott gives:

> Suppose that I reach out and catch a passing cricket ball. The next thing along in the ball's direction of motion was a solid brick wall. Beyond that was a window. Did my action prevent the ball hitting the window? (Did it cause the ball to not hit the window?) Nearly everyone's initial intuition is, "No, because it wouldn't have hit the window irrespective of whether you had acted or not." To this I say, "If the wall had not been there, and I had not acted, the ball would have hit the window. So between us—me and the wall—we prevented the ball hitting the window. Which one of us prevented the ball hitting the window—me or the wall (or both together)?" And nearly everyone then retracts his initial intuition and says, "Well, it must have been your action that did it—the wall clearly contributed nothing." (McDermott 1995, 525)

McDermott himself endorses the revised judgment of the majority that he reports.

Nevertheless, I do not share the intuitions of McDermott (and of "nearly everyone" he asked) in the ball catching example. Stopping of the ball before reaching the wall cannot be said to have prevented the window's breaking, since the window was not in any real danger of being broken anyway, thanks to the presence of the solid wall. Imagine, if you like, that in front of the window was a huge military tank, rather than the brick wall, situated to protect the window from breaking. Then the ball catcher can hardly be given credit for preventing the window from breaking by the ball.[1]

John Collins gives a similar example of preemptive prevention:

> As the ball flew toward us, I leapt to my left to catch it. But it was you, reacting more rapidly than I, who caught the ball just

[1] If, instead of the ball, an ultra-piercing bullet was fired towards the window, which could penetrate the tank and shatter the window, and our catcher stopped *that*, then he would have done some real preventing.

> in front of the point at which my hand was poised. Fortunate for us that you made the catch. The ball was headed on a course that, unimpeded, would have taken it through the glass window of a nearby building. Your catch prevented the window from being broken.
>
> Or did it? Had you not made the catch, I would have caught the ball instead. My leaping to catch the ball made your catch redundant. Given my presence, the ball was never going to hit the window. (Collins 2000, 223)

Collins disagrees (as I do) with McDermott's judgment in McDermott's example, but thinks that his own example is different. In his example, the person who caught the ball did prevent the window's breaking:

> If neither of us had reached for the ball, then the ball would have hit the window. So between us—you and me—we prevented the ball from hitting the window. Which one of us prevented the ball from hitting the window—you or I (or both of us together)? Well, clearly it must have been you, for it was you and not I who made the catch. I contributed nothing. (Collins 2000, 223-224)

According to Collins, in McDermott's example, "The presence of the wall really does seem to make your catch irrelevant." (Collins 2000, 224)

Both McDermott and Collins think that, in their own respective examples, the ball catcher is the preventer of the window's breaking. (McDermott: "it must have been your action that did it—the wall clearly contributed nothing"; Collins "it was you and not I who made the catch. I contributed nothing.") Be that as it may, I think our shadow case is somewhat different from the two authors' examples. A more closely analogous scenario to our shadow case would be if there were two parallel solid walls, each sufficient, by itself, to stop the ball from reaching the window. The ball hits one of the walls, call it wall$_A$, and is stopped by it; and the other wall, call it wall$_B$, contributes nothing. On this analogy, wall$_A$ is clearly what did the preventing of the window's breaking, just as the opaque object A prevented the surface *S*'s getting lit; while wall$_B$ is a backup preventer of the window's breaking, just as object B is a backup preventer of the surface *S*'s getting lit. What is important from my point of view is that, whether we regard our shadow scenario as a case of preemptive prevention—taking e_{pl} as the effect-event—or as an ordinary causation of the emergence of the shadow S—taking e_{fs} as the effect-event—our judgments about what is causally responsible for the effect in question do not change.

4.

There is no general agreement, however, that cases of prevention and preemptive prevention, like the unbroken window examples above, are cases of causation. The so-called cases of "negative causation" or "absence causation"—such as preventions, omissions, lacks and the like—are puzzling for theories of causation. There are philosophers taking opposing sides on the issue of whether absence causation should be regarded as genuine causation or should be treated as pseudo or "quasi" causation.[2] Some philosophers of causation are inclined to take at least some cases of prevention and omission as legitimate cases of causation, and the challenge for them is to pin down what distinguishes such cases from those absences which should not count as instances of causation. Ordinary intuitions also tend to take some absences as causal and some others not so. For example, when we say, "The driver's failing to see the warning sign on the road caused this fatal traffic accident," we seem to be attributing the cause to an absence: the driver's not noticing the sign. And when we say, "This fatal traffic accident caused him not to make it to the party," we seem to be referring to an absence as the effect, i.e. him not making it to the party. Sometimes both the cause and the effect are taken to be negative events as in, "Lack of sufficient lighting on the road caused the driver to miss the road sign." In still other examples of absences, the alleged cause and the alleged effect fail to compose a causal claim: "Nobody's dropping a bomb on the North Korean leader caused him not to die."

There are well known problems with taking absences as causes or effects, one of which is that it allows too many things to be causes or effects. For

[2] See, for example, the debate between Dowe and Schaffer (Dowe 2004; Schaffer 2004). It is not my purpose in this paper to take a position on the *general* issue of whether absences have causal efficacy or not.

example, my not walking on the surface of the planet Mercury right now is a cause of my writing this paper right now (if I were walking on the surface of Mercury at this moment, I couldn't be writing this paper). And my writing this paper right now is a cause of my not being on vacation in Hong Kong (or any other city) right now.

Now, it seems plausible to think of a shadow as absence of (some amount of) light from a surface. Thus our shadow set-up in the diagram can be viewed as involving a case of prevention—prevention of light from striking the relevant surface. Those philosophers who think that (at least some) preventions are not cases of causation would demand a justification for why A's prevention of light from reaching *S*'s surface should be described as A's *causing* *S*, which is how I described it above. Let me first point out a difference between our shadow case and the typical cases of prevention such as the ones in McDermott's and Collins's examples above. When the ball headed straight towards the window was caught before it reached the window, there occurred no change in the window's physical appearance or properties: it was unbroken *before* the ball was caught and remained unbroken *after* the ball was caught. Not so in our shadow example. For one thing, when A (and B) were put there and the shadow *S* was formed due to the blockage of light by A, the area occupied by *S* on the ground started to become cooler, due to the photons being prevented to transfer energy to that area. So, there did occur a change in the world in the vicinity of *S* in terms of temperature drop on *S*'s region compared to its surroundings. Moreover, when light was prevented by A, the contour lines of *S*, which were not there before A (and B) were placed there, emerged on the ground. There were other changes too, of course, brought about by the presence of *S* on the ground, such as the darkness observed on the ground by an observer standing near *S*.

There were no such changes in the window whose breaking was prevented by the successful catch of the ball. This is the big difference between our scenario and typical prevention and other absence cases: prevention of light by A has observable impacts on the world. Hence someone who thinks that preventions are not causes because they do not create relevant kinds of changes in the world, need not view shadows as "passive preventions" in the same way. My view is that shadows have *causes*; they are caused by light blockers and the presence of a ground, screen or something of that sort on which the shadow is projected. (Without something for a shadow to project itself on, we only have an umbra, which is not a shadow but a dark region in space.) And shadows certainly produce *effects* which are all too familiar: you can cool off on a hot day in the shadow of a tree, you can take a photo of a shadow, some shadows can be scary or funny, solar and lunar eclipses are exciting for us, etc.[3]

But, if someone were to insist that causal talk involving shadows is objectionable on the grounds that it involves absence causation, let me point out that we could pose the puzzle of section 1 without talking about shadows at all. In this way we can turn our scenario into one involving "presence causation" instead of absence causation. For example, instead of taking as our effect the emergence of shadow *S*, we could take it to be the presence of the event of cooling of the region *S*.[4] In that case our puzzle becomes: *What is causing the temperature drop* in region *S*—A or B? My answer would be the same as before: A is causing it and B is a preempted backup cause of it.

5.

Another interesting feature of the situation in the shadow diagram is that it seems to pose a problem for David Lewis's well-known counterfactual analysis of causation (Lewis 1973). Although, as I argued, the presence of A is causally responsible for the shadow *S*, we do not have a series of actual events running from A to *S* that constitute a chain of counterfactually

[3] Roy Sorensen is another author who thinks "shadow" is a causal concept, i.e. shadows are both caused by and cause things. See (Sorensen 2008, 9, 12, 18, 192).

[4] This strategy is similar to a strategy of replacing absences with presences described by Schaffer: "given that the gardener napped and my flowers wilted, 'The gardener's not watering my flowers caused my flowers not to blossom', is to be interpreted as: the gardener's napping rather than watering my flowers caused my flowers to wilt rather than blossom" (Schaffer 2005, 301). So, we can restate our claim regarding the shadow case as: Light's being blocked by A caused the cooling of the surface *S*.

dependent events from A to *S*, which Lewis's analysis requires. The presence of B blocks completion of such a chain. Take, for example, the events:

d_1: the presence of the dark region between A and B

d_2: the presence of the dark region to the right of B,

and consider the counterfactuals:

If A had not been present, then d_1 would not have occurred

If d_1 had not occurred, then d_2 would not have occurred

If d_2 had not occurred, then *S* would not have formed.

These counterfactuals fail to yield a chain of counterfactually dependent events in Lewis's sense, because the second counterfactual is false: even if d_1 had not occurred, d_2 would still have occurred thanks to the presence of B.[5]

Hence the situation in the diagram poses a counterexample to Lewis's 1973 analysis of causation. And this case does not seem assimilable to the other problematic cases for that analysis, which Lewis tried to deal with by emending his original 1973 account in his 1986 "Postscript to 'Causation'" (Lewis 1986). Lewis's dissatisfaction with some of his emendations in that "Postscript" led him to offer a new counterfactual theory in 2000 (Lewis 2000). This improved theory accounts for causation in terms of the notion of influence, which is defined by Lewis as follows:

> Where *C* and *E* are distinct actual events, let us say that *C influences E* if and only if there is a substantial range C_1, C_2 ... of different not-too-distant alterations of *C* (including the actual alteration of *C*) and there is a range E_1, E_2 ... of alterations of *E*, at least some of which differ, such that if C_1 had occurred, E_1 would have occurred, and if C_2 had occurred, E_2 would have occurred, and so on. (Lewis 2000, 190)

[5] The falsehood of the second counterfactual also follows from Lewis's possible-world semantics for counterfactuals: some world where d_1 does not occur and d_2 does is closer to the actual world than any world where both d_1 and d_2 fail to occur.

An event *C*, then, is a cause of an event *E* if and only if *C* influences *E*, or there is an ancestral of influence from *C* to *E*.

As an illustration of how the influence theory works, let us look at an example of how this theory is supposed to take care of trumping preemption cases, which are among the most challenging cases of causation to deal with by a counterfactual approach. An example of trumping preemption was given by Jonathan Schaffer:

> Imagine that … the major and the sergeant stand before the corporal, both shout "Charge!" at the same time, and the corporal decides to charge.[…] Orders from higher-ranking soldiers trump those of lower rank. I hope you agree that the major's order, and not the sergeant's, causes the corporal's decision to charge …. (Schaffer 2000, 175)

Lewis thinks that his improved theory can handle Schaffer's example. According to the new criteria Lewis added, first we imagine altering the trumping event while keeping the trumped event the same, and see if there would be any change in the effect. Thus suppose the major shouted "Take cover!", instead, while the sergeant ordered "Charge!". The soldiers, who hear both commands simultaneously, would have taken cover. Secondly, we imagine altering the trumped factor while keeping the trumping factor the same, and see if the effect would be any different. Suppose the major shouted "Charge!" while the sergeant shouted "Take cover!". The soldiers would have charged. Thus in the first case there would be a change in the effect, whereas in the second case there would be no change in the effect. Therefore we can conclude that it is the major's shouting, and not the sergeant's that is a cause of the soldiers' charging, according to Lewis.

But the influence approach would produce undesired results in our case. In our example, suppose we altered the height of A, say made it higher, while we kept B unaltered. The effect *S* would change—it would become a longer shadow. (A similar effect would ensue if we moved A towards B instead of increasing its height.) Secondly, suppose we increased the height of B while A remained fixed. The effect *S would* change again—it would become a longer shadow. (A similar effect would ensue if we moved B to the right instead of increasing its height.) In other words, there is a range of alterations that can be made on A or on B, such that the corresponding range of alterations on S counterfactually depend on the alterations on A or on B. Thus, Lewis must conclude that not only the presence of A but also the presence of B influences S.[6] Then both A and B are independently causes of S, which makes A and B overdetermining causes of S on Lewis's influence theory. But this is contrary to our verdict above that *only* A is a cause of S, as B is preempted by A from causally connecting to S.[7]

References

Collins, John. 2000. "Preemptive Prevention." *Journal of Philosophy* 97 (4): 223–234. https://doi.org/10.2307/2678391

Dowe, Phil. 2004. "Causes are Physically Connected to Their Effects: Why Preventers and Omissions are not Causes." In *Contemporary Debates in Philosophy of Science*, edited by Christopher Hitchcock, 189–196. Malden, Oxford and Carlton: Blackwell Publishing.

Lewis, David. 1973. "Causation." *Journal of Philosophy* 70 (17): 556–567. https://doi.org/10.2307/2025310

Lewis, David. 1986. Postscript to "Causation". In *Philosophical Papers*, vol. 2, edited by David Lewis, 173–213. New York: Oxford University Press.

Lewis, David. 2000. "Causation as Influence." *Journal of Philosophy* 97 (4): 182–197. https://doi.org/10.2307/2678389

McDermott, Michael. 1995. "Redundant Causation." *British Journal for the Philosophy of Science* 46 (4): 523–544. https://doi.org/10.1093/bjps/46.4.523

Schaffer, Jonathan. 2000. "Trumping Preemption." *Journal of Philosophy* 97 (4): 165–181. https://doi.org/10.2307/2678388

Schaffer, Jonathan. 2004. "Causes need not be Physically Connected to Their Effects: The Case for Negative Causation." In *Contemporary Debates in Philosophy of Science*, edited by Christopher Hitchcock, 197–216. Malden, Oxford and Carlton: Blackwell Publishing.

Schaffer, Jonathan. 2005. "Contrastive Causation." *Philosophical Review* 114, 297-328. https://doi.org/10.1215/00318108-114-3-327

Sorensen, Roy. 2008. *Seeing Dark Things: The Philosophy of Shadows*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195326574.001.0001

[6] This means that we do not have a case of trumping causation in our example: neither A nor B "trumps" the other.

[7] I thank Istvan Arányosi and an anonymous referee for their useful comments.

Organon F 26 (2) 2019: 298–300
https://doi.org/10.31577/orgf.2019.26207

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

DISCUSSION NOTE

Contradiction of Modal Modification

Miloš Kosterec*

The theory of property modification studies the logic and semantics of such terms as *fake banknote*, *former president*, and *skilled surgeon*. Terms like *fake, former,* and *skilled* (among many others) are called property modifiers. In general, a property modifier combines with a property to make a new property. Supposedly, there are four main types of property modifiers: *intersective, subsective, privative* and *modal*. One way to model the semantic properties of a modifier is to specify the particular type of entailment it appears in. For example, in the case of an intersective modifier, if something is a *grey* elephant, we know that it is grey and that it is an elephant. Consider a subsective modifier like *skilled*. If we know that somebody is a skilled surgeon, then we know that he or she is a surgeon. Now consider a privative modifier like *fake*. If we know that something is a fake banknote, then we know that it is not a banknote.

Here, I discuss the specification as well as the provided explication of modal modifiers via entailments. I demonstrate that both the specification and the explication are contradictory. First, modal modifiers are specified as follows:

> The unique feature, however, that modal modifiers have is that they oscillate between being subsective and being privative. So if the premise is that *a* is an *alleged terrorist*, say, then it is logically

* Institute of Philosophy of the Slovak Academy of Sciences
✐ Institute of Philosophy, Slovak Academy of Sciences, Klemensova 19, 813 64 Bratislava, Slovak Republic
✉ milos.kosterec@gmail.com

> possible that *a* be a terrorist and it is logically possible that *a* not be a terrorist. (Jespersen and Carrara 2013, 563)

(Jespersen 2015) utilizes Transparent Intensional Logic when formally specifying the entailments that involve modal modification:

> *Modal.* $\lambda w \lambda t\ [[{}^0M_m\ {}^0F]\ {}^0a]$ entails
> $\lambda w \lambda t\ [{}^0\lambda w'\ [{}^0\lambda t'\ [[[{}^0M_m\ {}^0F]_{wt}\ {}^0a] \rightarrow [{}^0F_{w't'}\ {}^0a]]]$
> $\wedge\ {}^0\lambda w''\ [{}^0\lambda t''\ [[[{}^0M_m\ {}^0F]_{wt}\ {}^0a] \rightarrow [{}^0Non\ {}^0F_{w''t''}\ {}^0a]]]]$
> (e.g. an *alleged* assassin is maybe an assassin). (Jespersen 2015, 336–37)

The elimination rule for modal modifier M_m applied to property f is then stated as follows (see Jespersen and Primiero 2013, 104):

$$\frac{[[M_m\ f]_{wt}\ x]}{\lambda w'[\lambda t'[[[M_m\ f]_{wt}\ x] \rightarrow [f_{w't'}\ x]]] \wedge \lambda w''[\lambda t''[[[M_m\ f]_{wt}\ x] \rightarrow \neg[f_{w''t''}\ x]]]}$$

Now consider *alleged discoverers of the highest prime number* as an example of modal modification. The highest prime number cannot exist. Therefore, there are no such discoverers. This does not mean, however, that somebody, say Kurt Gödel, could not be alleged to be among such discoverers. But then, following both the informal and the formal specification of modal modifiers, it should be logically possible, i.e. there should be a possible world in which Kurt Gödel is one of the discoverers of the highest prime number. But surely there is no such world, because there cannot be such a number. Therefore, the general specification of modal modifiers leads to contradiction when applied to the data.

This generalizes to every use of modal modifiers with regard to properties that cannot be instantiated. Such a property may be the intension, e.g., of a contradictory property concept such as *married bachelor*. The contradiction need not always be present in the concept, however. Consider the property concept *the necessarily empty property*. In general, if P stands for the necessarily empty property, M is a modal modifier, and k is an individual, then $[M\ P](\mathrm{k})$ leads to contradiction according to both the informal and the formal specification of modal modifiers. We can block the contradiction by stipulating that modal modifiers ought to be applied only to possibly non-empty properties. Such a stipulation is merely ad-hoc, however.

Therefore, there is a need for the non-contradictory specification of modal modification. Perhaps switching from logical to epistemological possibility is in order, at least in the case of modal modifiers. (Jespersen and Primiero 2013) also seem to be suggesting non-factive approach to modal modifiers as a viable route of investigation:

> The *actual truth* of [MmF] a entails that one of two *possibilities* is realized: a being an F; a not being an F. Thus there is a striking similarity between *modal modifiers* and *non-factive attitudes.* (Jespersen and Primiero 2013, 98)
>
> The link between modal modifiers and non-factive attitudes probably runs deeper than we let on in the present paper. [...] list of 'plain nonsubsective' (in effect, modal) modifiers/adjectives: *potential*, *alleged*, *arguable*, *likely*, *predicted*, *putative*, *questionable*, *disputed*. With the exception of *potential*, they all have something attitudinal about them. And all of those attitudes are nonfactive. A bold hypothesis would be that almost all modal modifiers are parasitic on non-factive attitudes. (Jespersen and Primiero 2013, 98, footnote 10)

Funding

This paper was supported by VEGA Grant No. 2/0117/19 *Logic, epistemology and metaphysics of fiction.*

References

Jespersen, Bjorn, and Massimiliano Carrara. 2013. "A New Logic of Technical Malfunction." *Studia Logica* 101 (3): 547–81. https://doi.org/10.1007/s11225-012-9397-8

Jespersen, Bjorn, and Giuseppe Primiero. 2013. "Alleged Assassin: Realist and Constructivist Semantics for Modal Modification." In *Logic, Language, and Computation*, edited by Guram Bazhanishvilli, Sebastian Löbner, Vincenzo Marra, and Frank Richter, 94–114. Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-36976-6_8

Jespersen, Bjorn. 2015. "Structured Lexical Concepts, Property Modifiers, and Transparent Intensional Logic." *Philosophical Studies* 172 (2): 321–45. https://doi.org/10.1007/s11098-014-0305-0

Organon F 26 (2) 2019: 301–305
https://doi.org/10.31577/orgf.2019.26208

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

BOOK REVIEW

Martin Smith: *Between Probability and Certainty: What Justifies Belief*

Oxford University Press, 2016, 213 pages

Shih-Hsun Chen*

In the book *Between Probability and Certainty: What Justifies Belief*, Martin Smith provides his normic theory of justification (NTJ) in contrast to the risk minimization conception (RMC) which is the prevailing view of epistemic justification. In general, it is not necessary to claim that a justified belief implies this belief is true and it seems that people are accustomed to using the probability point of view to determine the status of justification of a belief, which is the higher the probability of a belief being true, the more justification we give to this belief. However, Smith tries to provide another option for us to deal with this "uncertainty situation." In Chapters 1–6, Smith develops his theory and compares it to RMC in various aspects of justification—explanation, normalcy, and the comparative; in the last three chapters, Smith gives some formal and technical results in his theory. In this book review, I present the main argument of the book by means of three examples (the lottery case, the laptop case, and the catered case) provided in this book and one example that I give in the conclusion which points out some possible insufficiencies of Smith's theory.

According to RMC, a belief will not be justified unless the probability of this belief being true is high enough. It seems that RMC fits the general use of probability in our ordinary life—a high probability of occurrence provides a good reason to believe that it will really happen, similar to the situation where, after hearing the weather forecast informing that there is a 90% chance of rain, I take an umbrella with me if I go outside.

But, problems may occur when applying RMC in the following case. Suppose I hold a single ticket in a fair lottery of one million tickets and I know one of

* University of West Bohemia
✐ Department of Philosophy, Faculty of Arts, University of West Bohemia, Sedláčková 19, 306 14 Plzeň, Czech Republic
✉ bb9124037@hotmail.com

the tickets will win the lottery. By some simple calculation I know the odds of my ticket's losing are 99.9999%. Suppose that 99% is high enough to be a threshold to determine the justified status of a given belief, the belief "my ticket will lose" is justified. Furthermore, not only for my ticket, but the probability of each of all the other tickets losing is 99.9999%.

Now, if we accept multiple premise closure, we are faced with a paradox. According to multiple premise closure, if one has justification for believing each of all premises, and these premises together deductively entail a conclusion, then one has justification for believing the conclusion (p. 6). By multiple premise closure, we can conclude that the belief "no ticket will win" is justified. According to the setting, we know "one ticket will win the lottery" (so this belief is justified), hence we arrive at an awkward situation that the belief "no ticket will win, and one ticket will win the lottery" is justified. This is called the "lottery paradox." In order to avoid the lottery paradox, there are two options—we can either deny multiple premise closure or the idea that "my ticket will lose" is justified. Smith chooses the latter.

Smith provides an alternative theory—the normic theory of justification (NTJ). According to NTJ, "one has justification for believing P iff P is normically supported by one's evidence" (p. 77), and "a body of evidence E normically supports a proposition P just in case the circumstance in which E is true and P is false requires more explanation than the circumstance in which E and P are both true" (p. 40).

As a result, "my ticket will not win the lottery" is not justified in NTJ as regardless of whether my ticket wins or not, it does not need more explanation. This does not mean that something abnormal will not happen in lottery cases, such as someone cheated in this lottery; rather, it means that when we accept the probability of my ticket's losing is 99.9999%, we also accept that "my ticket will win" may still happen in spite of its low probability. In relation to this view, regardless of whether my ticket wins or not, we do not need extra explanation since the probabilistic evidence has explained this. Of course, we will still feel surprised when something with very low probability happens; we may even think there must be something happening which is unknown to us which has led to this result and we need some explanation about it in addition to the probabilistic evidence. Smith calls this "for all intents and purposes' normically supported."

Let us consider another example from this book that illustrates what a justified belief is like in NTJ. Suppose I have set my laptop to turn on with a randomly generated background which is set to be one out of one million values red and the remaining 999,999 values blue. One day I go to a library desk and turn on my laptop, and before it turns on, I see my friend, Bruce, who is already working on his laptop, and I go to say hello. Upon arrival at his desk, I see his laptop showing a blue background. In this laptop's case, "Bruce's laptop is displaying a blue background" gets normic support by its evidence but "My laptop is displaying a blue background" does not. If "Bruce's laptop is displaying a blue background" is not true, there must be some explanation such as a strange optical illusion or color blindness which is unknown to me. But if "My laptop is displaying a blue background" is not true, we need no extra explanation despite its extremely low probability. Smith said: "If one's belief turns out to be false, then the error has to be explicable in terms of disobliging environmental conditions, deceit, cognitive or perceptual malfunction, etc. In short, the error must be attributable to mitigating circumstances of some kind and thus excusable, after a fashion" (p. 41).

Smith provides NTJ as a new framework to understand justification by requiring more explanation if the justified belief turns out to be false. Although this theory has merit, such as it is consistent with multiple premise closure and it can solve the lottery paradox, if we accept it, we must accept that some beliefs that are unlikely to happen are justified. The following catered case illustrates this situation.

Suppose I am holding a large dinner party to which I've invited 100 guests (denoted by guest-1, guest-2, …, guest-100), and all guests have replied saying that they will attend. Suppose that I know all the invited guests are honest, trustworthy and well-meaning and I have no reason to suspect that any of them won't attend (p. 72). In this case, for any n in 1–100, "guest-n will attend my party" is justified since if guest-n does not show up, based on the evidence, there must be some explanation such as a family emergency, car accident…; by multiple premise closure, we will have that "all guests will attend my party" is justified. Despite the fact that all the guests are trustworthy and if someone does not show up, there must be some explanation which is attributable to mitigating circumstances of some kind, it is still hard to believe that all 100 guests will attend my party, given the real past party experience.

The party case indicates an important issue: are justified beliefs suitable to be the premises of our practical reasoning? To illustrate clearly, let us modify the party case. Suppose that my dinner party is to be catered for and I have a huge bet with someone about whether every guest will come to my party.

Now if I tell the caterers to prepare for 100 people and someone does not show up, I will lose all my money and even go to jail. As I understand that every guest is honest and trustworthy, should I tell the caterers to prepare for 100 people? Smith thinks that "all 100 guests will attend my party" is justified, but it is irrational, based on its high risk, to act upon this belief.

Smith provides another two theories of justification relative to normic conception: threshold normic theory of justification and the interest relative threshold normic theory. The threshold normic theory of justification shows that "one has justification for believing P iff the degree to which one's evidence normically supports P is greater than a threshold t, which can be variable and/or vague" (p. 99) and "the interest relative threshold normic theory shows that to claim that the value of the threshold t is to be determined in part by one's practical interests" (p. 100). Under these two theories, we can adjust the value of threshold t with the actual situation to avoid running a very high risk; hence, the belief "everyone will attend" is justified but is not high enough to meet our practical interests.

Now, we can distinguish two senses of justification: epistemic sense and practical sense. Normic theory of justification meets the former and threshold normic theory meets the latter. Returning to the party case, in order to avoid a very high risk, we can raise the value of t (by some practical interests) to check whether "guest-n will attend my party" is normically supported and in this extreme case— I will lose all my money and even may go to jail if I tell the caterer to prepare for 100 guests and someone does not show up—maybe "guest-n will attend my party" is not justified for every n. In addition to NTJ, RMC must deal with the same issue: are justified beliefs suitable to be the premises of our practical reasoning? For example, in the lottery case, if I already know the probability of one ticket winning the lottery is extremely low and the belief "the ticket I would buy will lose the lottery" is justified, then is it rational to buy a ticket?

So far, Smith's approach seems to be a promising framework for understanding what justification is; nevertheless, the core of NTJ, that is, the requirement for more explanation in mitigating circumstances and normalcy, is not particularly addressed in this book. Although NTC fits our intuitions about what normal is and has some good formal results, the lack of detailed accounts of normalcy makes it difficult to determine which situation needs more explanation than the others and which situation is more normal than the others. However, what bothers us so much in the catered party case is that it is normal that each guest will attend the party and it also seems normal that somebody will not show up to such a large private party.

Furthermore, it is difficult to understand the role of statistical evidence in Smith's theory. Suppose now we have E1: there is a 90% chance of rain tomorrow, E2: there is a 10% chance of rain tomorrow, and P: it will rain tomorrow. In light of NTJ, E1 does not normically support P, neither does E2 and therefore P will not be justified; hence P will not get more normic support (or more justification) from E1 than E2. Intuitively, we think that E1 will support P more than E2 does and I believe it is "normal" to think in this way; maybe this kind of support is not about the status of justification? What kind of support is this? The use of probabilistic expressions does not necessarily mean that we presuppose the occurrences are random. It is normal for my ticket to win the lottery in the most normal worlds, since there must be a ticket which wins the lottery, but it may be not normal in the most normal world that despite the fact that I studied hard, it turned out I failed some exam, based on the evidence (experience) showing that if I study hard, the probability of passing an exam is 90%. Smith should provide more analysis on this kind of evidence.

Organon F 26 (2) 2019: 306–310
https://doi.org/10.31577/orgf.2019.26209

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

BOOK REVIEW

Willard Van Orman Quine: *The Significance of the New Logic*

Translated and edited by Walter Carnielli,
Frederique Janssen-Lauret, and William Pickering
Cambridge: Cambridge University Press, 2018, xlvii + 168 pages

Ádám Tamas Tuboly*

Analytic philosophy is not filled with critical editions, with formerly unpublished archive materials that are edited by professionals, or with recently translated texts that were available previously only for a restricted circle of native-speaker scholars. Though there were some nice exceptions recently (as Gregory Frost-Arnold's transcription and edition of the famous Quine–Tarski–Carnap Harvard-discussions), it still counts as an important event in the profession if something like that appears. These hardly accessible materials are important for various reasons, but they are of utmost concern to anyone who is interested in the history of philosophy because without these what one might produce are philosophically motivated histories (in worst case fictions), while with their help historically supported philosophies could be produced as well.

The recent publication of Quine's *The Significance of the New Logic* is thus more than welcomed in the community. What is that we are dealing with now? Quine was invited to hold a seminar in São Paolo for a few months in 1942. After delivering his lectures in Portuguese, Quine left there his prepared notes and the manuscript appeared as *O Sentido da Nova Lógica* in 1944. It functioned as the major textbook for philosophers and logicians in Brazil for decades (p. viii–xii). This book—the second edition of which has appeared in 1996—has been translated and edited by Walter Carnielli, Frederique Janssen-Lauret and William Pickering and published by Cambridge University Press.

* Institute of Philosophy of the Hungarian Academy of Sciences
✐ Institute of Philosophy, Research Centre for the Humanities of the Hungarian Academy of Sciences, 4. Tóth Kálmán st., Budapest, 1097, Hungary
✉ tuboly.adam@btk.mta.hu

Supported by the MTA BTK Lendület *Morals and Science* Research Group and by the MTA *Premium Postdoctoral Scholarship*.

The book consists of four major parts. At first, there is a short informal editorial introduction summarizing the contents of the book, providing some Brazilian background and noting the editorial conventions used throughout the translation. It should be noted right at the beginning that the editors did an amazingly great and conscious job by providing explanatory notes and comparisons with Quine's other works. The second part is a longer historical-philosophical introductory essay by Janssen-Lauret (I will discuss it below) about Quine, his book and its significance. These introductions are followed by the actual translation and text of Quine's small (less than 150 pages) logic-book. The final section of the book is another translation: when Quine taught his seminar in São Paolo, he was invited to give a short summarizing-like lecture about the new logic and the United States. The short paper (12 pages), "The United States and the Revival of Logic," translated by the editors of this book, is the Appendix that is followed by a helpful list of editorial notes and a detailed index of names and subjects.

The reader is struck by the fact that many-many passages of Quine's book are just summaries or paraphrases of his back then recent two logic-textbooks that appeared in English, *Mathematical Logic* (1940) and *Elementary Logic* (1941). Though it surely made good sense for him to patch together the most valuable insights and methods of logic from previous materials in order to introduce the subject to an audience that starts from almost zero (especially in war-time when Quine did not have much time and energy to construct *entirely* new lectures), from our current point of view it makes the material a bit more usual or casual than especially revealing.

Quine's small textbook consists of an introduction and four parts. The first part is called "Theory of Composition" and it is basically a general introduction to the theorems and techniques of what is called recently propositional or sentential logic. Quine goes through all the connectives, their reduction, sentence formation and truth tables. It is quite understandable why this book was used frequently and widely in Brazil as the introductory text of logic: Quine's presentation is short, precise, explicit, and always points to the heart of the matter. Writing already three other books (the first one was his Ph.D. dissertation) on formal logic has its mark on this text. The next part is about the theory of quantification with the usual subjects of quantifiers, variables, their relation to truth, validity, proofs and implication. Part three is entitled "Identity and Existence" dealing furthermore with intensional contexts as well; finally, the fourth part is devoted to "Class, Relation, and Number", that is, to Quine's summary of his recent philosophy of mathematics.

A huge part of the text could be read as a simple introduction to logic that might be really interesting to historians of logic to see how notions, ideas, techniques and presentations evolved around the 1940s. There are certain passages, however, that might have further significance. In part two, Quine discusses, for example, the practical application of the theory of quantification (based on his less-known paper from 1939, "Relations and Reasons"), and argues that the new logic could be highly useful in the context of insurance. By translating natural language into logical form, reducing equivalent claims to simpler ones and then translating them back into natural language, clauses of insurance contracts could be simplified and shortened (pp. 78–79). This is a highly interesting form and mode of argument in favor of the new logic as reasoning about its application was mainly restricted to the natural sciences that time and even translations into natural languages (or as Quine said, "everyday language") was not a major concern of logicians.

In "Identity and Existence" Quine discussed many such ideas that became definite for him in the forthcoming years, and in cases, decades. We find here a detailed argumentation of why intensional contexts do not obey the rules that govern extensional contexts, how purely and non-purely designative occurrences influence the questions of identity, and in general, how meaning is to be approached with regard analyticity and synonymy. Furthermore, Quine also talks a lot about quantification, values and existence, relating Russell's theory of descriptions to the idea that the burden of ontological commitments is related to values and not to the use of names (as they are always eliminable). The importance of this part (§§32–41) could not be overestimated as Quine devoted much of his energy to discuss these questions in the forthcoming years. We should be thus more than thankful to have this text translated finally into English as the mark of Quine's transitory phase during the war, after his appearance as a logician and before his return as the leading philosopher of the States.

Quine knew the significance of these passages as they were noted and emphasized in his correspondence with Rudolf Carnap. Nonetheless, our happiness has certain limits and bitterness since almost the entire part of the book about these questions was translated into English by Quine already in 1943; it became the famous "Notes on Existence and Necessity" paper. While there are, of course, certain differences, omissions and changes between the original Portuguese text and the English article, and all of these are noted both in Janssen-Lauret's introduction (pp. xxxiv–xl) and in the editorial notes (pp. 159–161), these seem to be rather minor developments and corrections to the details than major ruptures in Quine's position.

The appendix to the volume, the translation of Quine's single lecture about the United States and the status of logic, could have been an important one as well. Nonetheless, almost four pages from the twelve are reproductions from Quine's introduction to his Portuguese textbook. The other materials in the lecture—however short, rudimentary and sketchy they might be—are more interesting. Quine notes, for example, "[t]he questions of the foundations of logic, like those of any other science, cannot be answered within psychology itself (according to some authors) without our falling into an infinite regress. The problem of avoiding this regress, if indeed it exists—or of explaining why it doesn't exist, in the negative case—belongs to philosophy rather than to any of the natural sciences" (p. 146). While obviously, Quine does not formulate explicitly his commitment to the famous thesis of his later paper about naturalized epistemology, his highly cautious formulations ("if indeed," "according to some"), also do not testify the claim that he rejected the naturalization of epistemology through psychology. Be as it may be, this is an interesting note (especially in the context of presenting the nature and results of modern logic), but this is not discussed further by Quine, or by the editors in any of the introductions.

The strangely transitional character of the article is also shown by the remark that deduction plays a crucial point in the natural sciences as well (and not just in mathematics) since "[i]f we can derive from the hypothesis [...] a sentence which conflicts with established facts, then we know that we will have to abandon the hypothesis" (p. 147). This indeed sounds like a quite naïve formulation of falsification and shows no clear or hidden sign of the revisable character of logic and observational statements that became so important for Quine just within a few years. Perhaps both the above and this remark could be explained due to the nature of being a popular lecture and thus sacrificing certain ideas on the altar of understandability and dissemination became a risk that was worth to take. If that is true, then it is still interesting why these ideas and why in that form were mentioned but not elaborated on in more details.

Nonetheless, none of these topics are discussed in the introduction to the volume. We also do not get to know whom exactly invited Quine to Saõ Paolo and why was he invited at all. Maybe all traces of this have been lost, but that should have been important to note as well for historians. What is discussed in greater detail is Quine's relation to Carnap and the various aspects of that relationship. Janssen-Lauret shows—and that is a point that was not emphasized sufficiently in the literature—that "[u]nlike Carnap, Quine did not have cause to associate metaphysics with dangerous political authoritarianism. He always favored a modest, empirically informed ontology" (p. xix). Quine made this explicit at various points in the lecture (both with respect to natural sciences and to logic and mathematics), and that seems to be indeed an important diverging point from Carnap during the early 1940s.

Nonetheless, it is not at all evident, not even from this text, that Quine and Carnap meant the same thing by "metaphysics," especially with regard "dangerous political authoritarianism" (that would fit Otto Neurath's concerns much better). Quine's acceptance of metaphysics is especially interesting given his American milieu: in pragmatist circles, metaphysics was regarded by many (e.g. Dewey) as the expression of feelings, ways of lives, and an approach to regulate human conduct; metaphysics had a practical and pragmatic aspect. (Later in the 1950s it was Philipp Frank, another important logical empiricist, who emphasized the same pragmatic character of Carnap's critique of and approach to metaphysics as the expression of *Lebensgefühl*). How Quine ended up with the conception of metaphysics as ontology is a further historical question that might be important especially, as Janssen-Lauret emphasized (p. xxx), that the Portuguese book contains many arguments for ontology and ontic-commitment for the first time.

Quine's critique of Carnap's and in general the Vienna Circle's (alleged) conception of conventionalism as the empiricist approach to logic and mathematics was noted in the introduction (pp. xxxviii–xxxix) as well. In the book, Quine made quite explicit and sharp statements about the drawbacks of conventionalism and about his own stance toward the matter as he did a few years before in "Truth by Convention." The translation is thus indeed highly valuable for these passages (mainly on pp. 14–15, 152–153 as this entry is missing from the index).

I have talked only about what is missing from the general and long introductory essay; but it should be noted as well that what is there is highly informative, well-structured and revealing about the book and Quine's context and influence in the history of analytic philosophy. The reviewer's concern shall be taken, thus, only to indicate that perhaps there is even more from a philosophical and historical point of view than was taken up by Janssen-Lauret. Thus, it may be the case that even if the book is not that much of a surprise and significant as it was envisioned before its English translation, we still have a nice material in our hands that deserves to be on our shelves as well. The perfectly edited pages and the highly personal cover of the book make it an even more appealing Cambridge volume.

Organon F 26 (2) 2019: 311–316
https://doi.org/10.31577/orgf.2019.26210

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

BOOK REVIEW

Lee McIntyre: *Post-Truth*

Cambridge, MA: MIT Press 2018, 240 pages

Pavol Hardoš*

Lee McIntyre's book *Post-Truth* (2018), part of the MIT Press' *Essential Knowledge* series, attempts the unenviable task of pinning down a vague, but very popular concept in our discourse. He settles on the understanding that post-truth denotes the notion of feelings being more accurate than facts, of believing something because it feels right. This also implies the potential for ideological domination by politically subverting the possibility of gathering facts about the real world. Interestingly, the implications of this latter half of his definition do not receive as much attention. Instead McIntyre focuses on the personal responsibility of epistemic agents to discover truth and the confluence of developments that made it so much harder for them.

The book's primary audience are lay people curious about the ongoing discursive practice of labeling lies and disinformation as *post-truth*. The book correctly reminds us that politically motivated denial of facts is not a creature of the current American electoral cycle. It offers a sweeping overview of why the phenomenon occurs—and why it appears to be everywhere today. McIntyre makes some very good points about the history and toxicity of science denialism, the nature of our motivated thinking, the development of the prestige press, the idea of objectivity in media, the fragmentary effects of social media information silos, and so on—though these are hardly novel, it is commendable to have them explained briefly and accessibly.

The book is ultimately unconvincing, however, not just because it appears to suffer with symptoms of what it diagnoses—post-truth errors of both fact

* Comenius University
✐ The Institute of European Studies and International Relations, Faculty of Social and Economic Sciences, Comenius University, Mlynské luhy 4, 821 05, Bratislava, Slovak Republic
✉ pavol.hardos@fses.uniba.sk

and interpretation (more on that later)—but because for a work that seeks to tackle an epistemological issue—even if in a popular vein—it does not really engage with the relevant literature on social epistemology. The book neglects the very essential epistemological questions any treatment of truth (post- or otherwise) needs tackling: what truth is and how do we come to believe it in the first place. Neither do we get a convincing account of why post-truth is a distinct phenomenon [for a recent skeptical take, see (Habgood-Coote 2018)], and not a moral panic, a conceptual muddle of lies, propaganda, and bullshit (in the Frankurtian sense), or merely a discursive shortcut for numerous disquieting social, political, and technological developments. Instead we get the by-now somewhat tired chapters on science denialism, cognitive biases, the decline of traditional media, the rise of social media and—with a surprising twist—the blameworthiness of post-modernism.

The errors of fact can be illustrated by the following examples. The chapter on cognitive biases discusses the backfire effect, the notion that corrective information can not only fail to register but make the recipient of the correction double down on the falsehood and believe it even more strongly. This effect, however, has famously failed to replicate (Wood and Porter 2016)—with the study's original authors co-authoring a further replicating study with a similar lack of results (Nyhan et al. 2017). This problem was known for almost a year before this book went to print yet is not acknowledged anywhere. It was almost as if this fact failed to register.

Another curious error can be found in the final chapter on combating post-truth and the need to strongly challenge lies and deceptions in a timely manner. Here the lesson starts with a parable that John Kerry failed to react strongly to lies during the 2004 presidential campaign and consequently "lost the election by a few thousand votes in Ohio" (p. 155). A cursory search for the results quickly reveals those 'few thousand' votes to be 118 thousand, or a margin of slightly more than 2%. (George W. Bush also won the popular vote by about 3 million, but let's not get inconvenient facts in the way of a good narrative.)

The errors of interpretation require a bit more space. Here his chapter about post-modernism is emblematic of the books' weaknesses. McIntyre's basic argument is that post-truth as a modern phenomenon was enabled by the developments in post-modern philosophy, which problematized the notion of objective truth as unideological and apolitical, wholly disconnected from the world of human power, interpretation, and values.

The chapter makes valid points about incongruous pronouncements from certain *science and technology studies* scholars. McIntyre shows many to have gone beyond circumspect critiques of the ways scientific findings or concepts come to be treated as facts into outright denial of facts: it is all ideology anyway. McIntyre makes a great deal out of the famous, heart-felt mea culpas from Bruno Latour (2004), one of the most famous scholars who talked about social construction of scientific facts, but who now wishes to restore the idea of scientific fact as something objectively true.

But McIntyre's argument is far from smooth. His primary argument follows the one in a paper by philosopher of science Robert Pennock (2010) about Phillip Johnson, the god-father of the Intelligent Design (ID) movement. Johnson consciously cited critical theory and relativism he had read about in law school as his operating principles for advancing his preferred creationist version of biological explanations. From here McIntyre makes a jump to other instances of science denial, such as climate science denialism, or anti-vaccination movements, for which the ID movement served as a blueprint. But the said blueprint consisted of examples of funding 'counter-research' and pushing their own 'experts' to create the illusion of controversy and debate, not from a relativistic deconstruction of scientific practices, a point McIntyre elides.

McIntyre further credits the Sokal hoax for bringing post-modern posturing into the mainstream but is unwilling to extend the blame for the fallout of this wider awareness, even though this is crucial for his argument elsewhere. Earlier he laments that these post-modern notions 'leaked' into wider consciousness and have been used unscrupulously beyond obscure academic journals. I am not saying we should be blaming Sokal too, for popularizing post-modern intellectual posturing, only that for McIntyre to be consistent in his belief that people are blameworthy for how their ideas are used (never mind what were their intentions), he must also lay blame at the feet of those who propagate such ideas, whatever their intentions.

But most of all, his treatment of 'post-modernism' as one of the sources of our current post-truth predicament seems more ideological than anything else. It is far too easy to blame an ill-defined, elusive concept such as post-modernism for post-truth. McIntyre echoes long-standing conservative obsessions with post-modernism (or "cultural Marxism" or "critical theory" in other, similar iterations) as a scourge of truth and beauty instead of what it really is: a set of divergent, theoretical propositions about knowledge in our society. Here he joins the narrative of the likes of Dennett, Pinker, Dawkins et al. who are at the forefront of the discursive efforts to straw-men all post-modern criticism of how science is done into a belief system committed to radical skepticism at best, or a relativizing incoherence, at worst.

No privileged elder statesmen of science and objective truth probably like vexing inquiries about their potential biases or about why their intellectual pronouncements go beyond their immediate expertise. Though to point out this vested interest would probably already reveal one as a post-modernist, too. Any recognition of the plurality of discourses and perspectives about the world would do that, yet this post-modern reflection on the lack of a monopolized control over meta-narratives does not commit one to a full-blown relativist standpoint.

Indeed, not all post-modernist constructivism in science is the enemy of the quest for truth—on the contrary, one cannot get to truth without realizing the extent of subjectivity when we ask research questions, build concepts, choose the tools, & model the world and how this—often unconscious—dealing with the world around us can color our perceptions of the world.

According to McIntyre's veritably post-factual treatment of Derrida and Foucault, they are radical sceptics, nihilists claiming it is all *only* about the text and/or power. However, they did not really deny the possibility of objective reality (cf. Prado 2006). Contrary to McIntyre's (especially) unfair portrayal of him, Foucault would probably not agree that professions of truth are "nothing more than a reflection of the political ideology of the person" making them (p. 126). Knowledge claims are not "just" assertions of authority, a "bullying tactic" used by the powerful (p. 126)—but it is important to realize that *they can be*. In search for truth we must be aware of this possibility and add this warning into our calculus of trust over particular claimants and their claims to authority. This is a profound insight that we credit Foucault and other scholars with. Without it our understanding of objective reality would be much poorer. We cannot be blind to the truth that knowledge claims are potentially *also* ideological. This is not necessarily a rejection of objective reality, this is a reminder of the warranted distrust towards those who have historically claimed to own the truth.

Claims to truth must be interrogated with an eye to the context in which they were made to spot any potential biases or alternative explanations. This is no truth-denying relativism but sound epistemic practice, one which is still far from being the norm. Espousing such commitment to skepticism over knowledge claims does not commit one to denialism. Only very uncircumspect or naïve people would make that conceptual jump, but McIntyre seems only too willing to push his readers to precisely such somersaults about postmodernism.

In the final analysis, McIntyre also offers advice on how to fight post-truth, but it is equally un-inspiring. He admonishes us to take responsibility over our personal epistemic practices: be skeptical, buy a quality newspaper now and then, fight the instinct for partisanship and confirmation bias—we can do it, it is our decision how we react to the world. There is no accounting of structural issues, institutions and their epistemic effects (cf. Rini 2017), or simply of how ridiculous it is to epistemically pull yourself by your bootstraps out of bullshit in the information environment he described in the previous chapters.

Thus, the biggest missed opportunity of the book is that, in our current environment ripe for educating the lay public about how we come to know and trust things as factual, it does not take social epistemology seriously enough—it completely neglects the discussion of testimony (e.g., Lackey 2008), reputation (e.g., Origgi 2017), and the individual and social norms, as well as institutions (e.g., Goldman 1999), that make knowing and believing the truth possible. Instead, apart from offering pop-science explanations, it seems intent on waging a clandestine ideological proxy war—right in the spirit of the times it purports to diagnose.

References

Goldman, Alvin. 1999. *Knowledge in a Social World*. Oxford: Oxford University Press. https://doi.org/10.1093/0198238207.001.0001

Habgood-Coote, Joshua. 2018. "Stop Talking about Fake News!" *Inquiry*. https://doi.org/10.1080/0020174X.2018.1508363

Lackey, Jennifer. 2008. *Learning from Words: Testimony as a Source of Knowledge*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199219162.001.0001

Latour, Bruno. 2004. "Why Has Critique Run Out of Steam? From Matters of Fact to Matters of Concern." *Critical Inquiry* 30 (2): 225–48. https://doi.org/10.1086/421123

Nyhan, Brendan, Ethan Porter, Jason Reifler, and Thomas Wood. 2017. "Taking Corrections Literally but Not Seriously? The Effects of Information on Factual Beliefs and Candidate Favorability." *SSRN*. Accessed September 1, 2018, https://ssrn.com/abstract=2995128

Origgi, Gloria. 2017. *Reputation: What It Is and Why It Matters*. Princeton, NJ: Princeton University Press.

Pennock, Robert. 2010. "The Postmodern Sin of Intelligent Design Creationism." *Science & Education* 19 (6–8): 757–78. https://doi.org/10.1007/s11191-010-9232-4

Prado, C.G. 2006. *Searle and Foucault on Truth*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511616020

Rini, Regina. 2017. "Fake News and Partisan Epistemology." *Kennedy Institute of Ethics Journal* 27 (S2): 43–64.

Wood, Thomas, and Ethan Porter. 2016. "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence." *SSRN*. Accessed September 1, 2018. https://doi.org/10.2139/ssrn.2819073