VOL. VI, NO. 39, NOVEMBER 2014

DISPUTATIO

INTERNATIONAL JOURNAL OF PHILOSOPHY

Articles

Critical Notice

Book Reviews

Centro de Filosofia da Universidade de Lisboa www.disputatio.com

Disputatio publishes first-rate articles and discussion notes on any aspects of analytical philosophy (broadly construed), written in English or Portuguese. Discussion notes need not be on a paper originally published in our journal. Articles of a purely exegetical or historical character will not be published.

All submissions to *Disputatio* are made by email to disputatio@campus.ul.pt. Please read the instructions on our site before submitting a paper. *Disputatio* requires authors to include a cover letter with their submission, which must contain all useful contact information, as well as the title of the submitted article, keywords and word count. Submissions must be either in English or Portuguese. A short but informative abstract (around 100 words) at the beginning of the paper is required, followed by 5 keywords.

All Unsolicited Contributions to *Disputatio* are triple-blind refereed: the names and institutional affiliations of authors are not revealed to the Editors, the editorial committee and editorial board, or to the referees. Without the prior permission of the Editors, referees and Board members will not show to other people material supplied to them for evaluation. All published submissions have been anonymously reviewed by at least two referees.

Submissions and email are to be sent to disputatio@campus.ul.pt, or to *Disputatio*, Centro de Filosofia da Universidade de Lisboa, Faculdade de Letras, Alameda da Universidade, 1600-214 Lisboa, Portugal. Publishers should send review copies to Teresa Marques at this address.

Founded in 1996, *Disputatio* was published by the Portuguese Philosophy Society until 2002. From 2002, it is published by the Philosophy Centre of the University of Lisbon. *Disputatio* is a non-profit publishing venture. From 2013, *Disputatio* is published only online, as an open access journal.

PUBLISHED BY

VNIVERSITAS Centro de Filosofia OLISIPONENSIS

SPONSORED BY

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Directores: João Branquinho e Teresa Marques. Publicação semestral. N.º de registo no ICS: 120449. NIPC: 154155470. Sede da redacção: Centro de Filosofia, Faculdade de Letras de Lisboa, Alameda da Universidade, 1600-214 Lisboa.

DISPUTATIO

INTERNATIONAL JOURNAL OF PHILOSOPHY

Vol. VI, No. 39, November 2014

Published by Centro de Filosofia da Universidade de Lisboa
ISSN: 0873 626X — Depósito legal n.º 106 333/96

Higher-Order Vagueness and Numbers of Distinct Modalities

Susanne Bobzien
University of Oxford

BIBLID [0873-626X (2014) 39; pp. 131-137]

Abstract
This paper shows that the following common assumption is false: that in modal-logical representations of higher-order vagueness, for there to be borderline cases to borderline cases ad infinitum, the number of possible distinct modalities in a modal system must be infinite.

Keywords
Vagueness, higher-order vagueness, modalities, modal logic, KT4.

There is a not uncommon misconception regarding the relation between higher-order vagueness and the number of distinct modalities in a modal system. It is this.

> (1) For a theory of higher-order vagueness to be useful towards the solution of the Sorites (by eliminating any detectable sharp boundary between non-borderline and borderline cases), it must permit the expression of *radical higher-order vagueness*, i.e. of borderline borderline … borderline cases, for any number n of iterations of 'borderline'. (2) An object a is borderline F or indeterminately F (IFa) precisely if it is not determinately F and not determinately not F ($\sim DFa \& \sim D \sim Fa$). (3) If higher-order vagueness is expressed by means of axiomatic — or other — systems of modal logic, the number of distinct modalities of the system must be infinite for it to be possible that there is radical higher-order vagueness.

In this paper we take issue with (3). (3) is usually rolled out as an objection to the claim that the modal system KT4 (or S4) may be suit-

Disputatio, Vol. VI, No. 39, November 2014
Received: 13/02/2014 Revised: 16/06/2014 Accepted: 19/07/2014

able to represent higher-order vagueness. Here is a recent example:

> If S4 (i.e. KT4) is the logic for absolute definiteness then there is only a finite number of modalities (in fact at most fourteen distinct modalities, see Chellas 1980, 149). Consequently, there cannot be borderline cases to borderline cases ad infinitum. (Åkerman and Greenough 2010: 287, n.37.) [1]

Evidently, this objection is not restricted to KT4. A *modality* is any sequence of the operators ~, $\Box$, $\Diamond$. Two modalities Φ, Ψ are *distinct* if and only if for some A $\Phi A \leftrightarrow \Psi A$ is not a theorem. So, if we confine ourselves to familiar systems of normal modal logics and add the fact that axiom **T** seems universally accepted for logics of vagueness, then KT and KTB would be *prima facie* suitable, since either has infinitely many distinct modalities. On the other hand, KT4, KT5, KT4G and $KT4G_c$ would each be unsuitable for expressing radical higher-order vagueness, since the number of their modalities is finite.

Why would anyone think this? Åkerman and Greenough don't give much away in the paper quoted: they seem to imply that for it to be possible for there to be borderline cases to borderline cases ad infinitum (i.e. radical higher-order vagueness) there need to be infinitely many distinct modalities. Let's make the plausible assumption that this is taken to be so because each order of borderlineness needs its own distinct modality, or set of distinct modalities. For there to be borderline cases, there needs to be at least one modality; for there — also — to be borderline borderline cases, there need to be at least two, etc. Why would anyone think *this*? It is safe to assume that the underlying assumption is that for there to be genuine higher-order borderline cases, the *extension* of the borderline borderline cases must differ from that of the singly borderline cases, that of the triply borderline cases must differ from that of the doubly and the singly borderline cases, etc.

In fact, (3) from above indicates a misunderstanding of the nature of genuine higher orders in higher-order vagueness. It is a mistake to think that the number of distinct modalities in a modal system S limits the number of possible higher orders. More specifically, theorems

[1] This argument is different from the objections against axiom 4 that Williamson raises (1994: 157-61) and which are followed up by Greenough 2005. For some rejoinders to those objections see Bobzien 2012: 194-200, 204-210.

expressing material equivalence between iterative formulas of different ranks[2] in a logic of vagueness (e.g. $DA \leftrightarrow D^2A$) do not eliminate genuine higher orders. Compare epistemic logic. Assume for the sake of argument that it is logically true in some epistemic logic that I know that *A* if and only if I know that I know that *A*. Then I would still have genuine second-order knowledge if 'I know that I know that *A*' is true. Exactly the same holds for higher orders of vagueness. In terms of modalized predicates, if DFx and D^nFx are co-extensional for any *n*; or if IFx and I^nFx are co-extensional for any *n*, either way, this does not preclude that there are *a* that are genuinely I^nF. Take, for example, an epistemic interpretation of IFa as '*a* is such that one can't tell that it is *F* and one can't tell that it is not *F*', or, for short, '*a* is such that one can't tell whether it is *F*'. Assuming compositionality (and the mirror axiom $IA \leftrightarrow I{\sim}A$), I^2Fa then stands for '*a* is such that one can't tell whether one can tell whether it is *F*'. Even if IFa and I^2Fa are extensionally equivalent, they clearly express two different things. It is one thing for someone to be unable to tell whether *Fa*, and another for someone to be unable to tell whether they are unable to tell whether *Fa*. The same holds for higher orders. In particular if *a* is such that one can't tell whether one can tell … (indefinite times) … whether one can tell whether it is *F*, then contrary to (3) there is radical higher-order vagueness.

It is not necessary to take an epistemic interpretation. Consider instead some semantic or ontic interpretation of the indeterminacy. For instance, interpret IFa as 'it is semantically indeterminate whether *Fa*'. Assuming compositionality (and the mirror axiom $IA \leftrightarrow I{\sim}A$), I^2Fa then stands for 'it is indeterminate whether it is indeterminate whether *Fa*'. Again, even if IFx and I^2Fx are extensionally equivalent, they clearly express two different things and contrary to (3) there is radical higher-order vagueness.

One purpose of a logic of vagueness (or indeterminacy or borderlineness) is to provide a representation of the — or certain — structural properties of vagueness (or indeterminacy or borderlineness). There is nothing inherent in the notions of determinacy or indeterminacy that prohibits co-extensionality of the determinate

[2] DA is of rank 1, D^nA of rank *n*, etc. For a recent formal definition of modal ranks (or modal degrees) see e.g. Carnielli and Pizzi 2009: 27-8.

and the determinately determinate, or of borderline cases and borderline borderline cases. ([~*DFa*&~*D*~*Fa*] & [~*D*[~*DFa*&~*D*~*Fa*] & ~*D*~[~*DFa*&~*D*~*Fa*]] is coherent in a system that contains PC, MP, N, **K** and **T**.) It is perfectly possible to have infinite orders of determinacy and of borderlineness with a finite number of distinct modalities.[3] Note also that it follows from, and for, Williamson's account of higher-order vagueness that, if in KT4 some *A* has second-order vagueness, it has vagueness at every order (Williamson 1999: 132-3, 136).[4]

We conclude by considering two retorts which are sometimes voiced. *Retort 1*: "Agreed, there can be infinite orders of determinacy and borderlineness with a finite number of distinct modalities; however, this can be achieved only at the expense of introducing detectable sharp boundaries between determinate cases and borderline cases." One can see how someone might get this idea by examining KT5 and KT4 and coming to the conclusion that neither is suitable for eliminating sharp boundaries. Given (2) and modal axioms **4** and **5**, KT5 provides, for a vague predicate F, only (i) determinate cases of *F*, (ii) determinate indeterminate cases of F and (iii) determinate cases of ~*F*. This suggests sharp borders into and out of the borderline zone. And Williamson (1999: 134) shows that with *his own* formal characterization of higher-order vagueness, system S5 is the weakest extension of KT that would permit vagueness and forbid higher-order vagueness. As for KT4, it may appear to lead to a

[3] This holds regardless of whether higher-order vagueness is defined (i) as '*A* is n^{th}-order vague if I^nA (and *F* is n^{th}-order vague if $\exists xI^nFx$)'; or (ii) with Williamson (1999: 132) as "[w]e have a first-order classification of states of affairs according to whether *A* or ~*A* holds. Vagueness in the first-order classification is first-order vagueness in *A*. [...] we have an $(n+1)^{th}$-order classification according to whether members of the *n*th-order classification definitely hold, definitely fail to hold or are borderline cases. Vagueness in the *n*th-order classification is *n*th-order vagueness in *A*"; or (iii) in any other way directly based on (2).

[4] In Williamson's account (see previous note), ~DD^nA&~*D*~D^nA with $n\geq 0$ is a sufficient condition for $(n+1)^{th}$ order vagueness. By $DA \leftrightarrow D^nA$ for $n\geq 1$ in KT4 we get (i) ~*DDA*&~*D*~*DA*→~DD^nA&~*D*~D^nA. We get (ii) ~*DDA*&~*D*~*DA*→~*DA*&~*D*~*A* by the KT4 theorems (iii) *DA*→*DDA* and (iv) ~*D*~*DA*→~*D*~*A*: (iii) together with the contraposition of (iv) provides *DA*∨*D*~*A*→*DDA*∨*D*~*DA*, which by contraposition and DeMorgan gives (ii). (ii) covers the case of $n=0$ and (i) covers the cases with $n>0$.

sharp border from the *n* times determinate cases (D^nF) to the *n* times borderline cases (I^nF) at the beginning of some assumed borderline zone and for indefinite *n*. However, in both cases the argument is *not* that the extensions of the borderline, and the borderline borderline, cases, etc., are co-extensive. Rather, for KT5 the argument is *that there is a sharp boundary* between the determinately determinate cases and the determinate borderline cases; and for KT4 it would be *that there is a sharp boundary* between the cases that are D^nF and the borderline cases that are I^nF. Thus, even though KT5 and KT4 may have been shown to be unsuitable for avoiding a sharp boundary, it has not been shown that this is so *because* the number of their distinct modalities is finite. More importantly, system KT4G_c or S4M, which adds axiom $\mathbf{G}_c$ ($\Box\Diamond A \rightarrow \Diamond\Box A$) to KT4, and which has only a measly eight distinct modalities, both preserves higher-order vagueness and complies with the intuition that there are no detectable sharp boundaries between borderline and non-borderline cases. In its determinacy version it has both $DA \leftrightarrow D^2A$ and $IA \leftrightarrow I^2A$ as theorems and thus introduces infinite orders of both determinacy and indeterminacy (or borderlineness). At the same time KT4G_c defines a logic of determinacy that has as one of its inherent features that no sharp boundary between the borderline cases and the non-borderline cases can be determined.[5]

Retort 2: "By *a* being borderline *F* we don't *just* mean $\sim DFa \& \sim D \sim Fa$. The borderline cases also have to be *between* the determinate cases." This is, of course, changing the rules halfway through the game. Instead of the standard modal account of borderlineness (2) from above, we now have something like this (we offer a charitable version), with *BLFa* for *a* is borderline *F*:

(4) *BLFa* if and only if [$\sim DFa \& \sim D \sim Fa$]&*a* is between the things that satisfy *DF* and the things that satisfy $D \sim F$.

(5) BL^2Fa if and only if [$\sim DBLFa \& \sim D \sim BLFa$]&*a* is between the things that satisfy *DBLF* and the things that satisfy $D \sim BLF$.

[5] Bobzien 2010 provides an extended argument for the compatibility of radical higher-order vagueness with axiom 4 and with the characteristic axiom of KT4G_c.

It is accounts of borderlineness along the lines of (4) and (5) which open the door for the so-called higher-order vagueness paradoxes.[6] We believe that such accounts and the ensuing presumed paradoxes are the result of a confusion between higher-order vagueness and the distribution of the objects of a Sorites series into extensionally non-overlapping categories.[7] But even with (4) and (5), the numbers of higher orders do not depend on the numbers of distinct modalities: with a sufficiently fine-grained Sorites series nothing prevents there from being more than, say, fourteen higher orders. In any event, we set out to show that, given (2), (3) is false; i.e. that, given (2), there cannot be "borderline cases to borderline cases ad infinitum", even with a finite number of distinct modalities such as in KT4. And this we have shown.[8]

Susanne Bobzien
All Souls College
University of Oxford
Oxford OX1 4AL, UK
susanne.bobzien@philosophy.ox.ac.uk

References

Åkerman, Jonas and Greenough, Patrick. 2010. Hold the Context Fixed: Vagueness Still Remains. In *Cuts and Clouds: Vagueness, its Nature, and its Logic*. Edited by Dietz, Richard and Moruzzi, Sebastian. Oxford: Oxford University Press, 275-88.

Bobzien, Susanne. 2010. Higher-order Vagueness, Radical Unclarity, and Absolute Agnosticism. *Philosophers' Imprint* 10: 1-30.

Bobzien, Susanne. 2012. If it's Clear, then it's Clear that it's Clear, or is it? — Higher-order Vagueness and the S4 Axiom. In *Episteme, etc.* Edited by Katerina Ierodiakonou and Benjamin Morison. Oxford: Oxford University Press, 189-212.

Bobzien, Susanne. 2013. Higher-order Vagueness and Borderline Nestings — a Persistent Confusion. *Analytic Philosophy* 54: 1-43.

[6] See Fara 2003: 196-200, Sainsbury 1991: 167-70, Shapiro 2005: 147-51, Wright 1992: 129-33, 137 and Greenough 2005: 182-3 for different versions of this type of presumed paradox.

[7] For a detailed account of this confusion see Bobzien 2013.

[8] Thanks to Nicholas Denyer and to an anonymous referee from *Disputatio* for helpful comments.

Carnielli, Walter and Pizzi, Claudio. 2009. *Modalities and Multimodalities*. New York: Springer.

Chellas, Brian F. 1980. *Modal Logic: An Introduction*. Cambridge: Cambridge University Press.

Fara, Delia Graff. 2003. Gap principles, penumbral consequence and infinitely higher-order vagueness. In *Liars and Heaps: New Essays on Paradox*. Edited by J. C. Beall.Oxford: Oxford University Press, 195-222. Originally published under the name 'Delia Graff'.

Greenough, Patrick. 2005. Contextualism about Vagueness and Higher-Order Vagueness. *Proceedings of the Aristotelian Society* (suppl) 105: 167-90.

Sainsbury, Mark. 1991. Is There Higher-Order Vagueness? *Philosophical Quarterly* 41: 167-82.

Shapiro, Stewart. 2005. Context, Conversation, and so-called 'Higher-Order Vagueness'. *Proceedings of the Aristotelian Society* (suppl) 105: 147-65.

Williamson, Timothy. 1994. *Vagueness*. London: Rougledge.

Williamson, Timothy. 1999. On the structure of higher-order vagueness. *Mind* 108: 127-142.

Wright, Crispin. 1992. Is Higher-Order Vagueness Coherent? *Analysis* 52: 129-39.

Right-Making, Reference, and Reduction

Michael Byron
Kent State University

BIBLID [0873-626X (2014) 39; pp. 139-145]

Abstract
The causal theory of reference (CTR) provides a well-articulated and widely-accepted account of the reference relation. On CTR the reference of a term is fixed by whatever property causally regulates the competent use of that term. CTR poses a metaethical challenge to realists by demanding an account of the properties that regulate the competent use of normative predicates. CTR might pose a challenge to ethical theorists as well. Long (2012) argues that CTR entails the falsity of any normative ethical theory. First-order theory attempts to specify what purely descriptive property is a fundamental right-making property (FRM). Long contends that the notion that the FRM causally regulates competent use of the predicate 'right' leads to a *reductio*. The failure of this argument is nevertheless instructive concerning a point at which ethics and metaethics overlap.

Keywords
Normative property, descriptive property, causal theory of reference, Jackson, Schroeder

Right-making, reference, and realism

The causal theory of reference (CTR) provides a well-articulated and widely-accepted account of the reference relation. On CTR the reference of a term is fixed by whatever property causally regulates the competent use of that term. CTR poses a metaethical challenge to realists by demanding an account of the properties that regulate the competent use of normative predicates.[1] For non-naturalistic realists,

[1] Since anti-realists generally deny that moral judgments involve predication, on their view the semantic value of moral judgments does not involve reference

Disputatio, Vol. VI, No. 39, November 2014
Received: 30/09/2013 Revised: 08/04/2014 Accepted: 30/04/2014

who assert that normative properties[2] are non-natural, the puzzle is to account for how non-natural properties might causally regulate anything. Non-naturalists like Shafer-Landau (2003) define normative properties in terms of non-identical concatenations of natural properties, but by denying identity such views threaten to deny that the reference of normative predicates is fixed. Non-naturalists could of course decline the challenge and jettison CTR. Naturalists, by contrast, regard normative properties as natural properties, and would seem to have an easier time accommodating CTR. Cornell realists like Sturgeon (1988) assert that normative properties are natural properties in their own right, and presumably such normative properties are available to play a causal role in reference. Reductive naturalists like Railton (1986) claim that normative properties are reducible to descriptive properties, and the reduction base might regulate competent use.

CTR might pose a challenge to first-order ethical theorists as well. Long (2012) argues that CTR entails the falsity of any normative ethical theory. First-order theory attempts to specify what purely descriptive property is a fundamental right-making property (FRM). Long contends (bracketing his discussion of the possibility of multiple FRM's) that the notion that the FRM causally regulates competent use of the predicate 'right' leads to a *reductio*. The argument relies on two assumptions, namely:

> A1. A purely descriptive property is a FRM only if the moral property of being right exists.
>
> A2. If the moral property of being right exists, then our predicate 'right' refers to it.

By CTR, if a property F causally regulates competent use of the predicate 'right', then 'right' rigidly designates F. By A1 and A2, 'right'

---

to properties.

[2] I follow the now fairly standard usage of Jackson (1998: 120-121), according to which a normative property is a property that may be ascribed by a normative predicate, and a descriptive property is a property that may be ascribed by a descriptive predicate. For an accessible discussion of Jackson's reductionism, see Streumer 2011.

refers to and thus rigidly designates the property of being right. It follows that the property of being right is identical to F, and Long claims that this consequence renders the explanation of rightness absurd. Ethical theory postulates a FRM in order to explain the property of being right, but according to Long it is absurd to think that one property might explain another when they are identical. Hence the *reductio* of the claim that the FRM causally regulates 'right'. And this conclusion poses a dilemma: either there is no FRM, contrary to ethics, or nothing causally regulates 'right', contrary to CTR.

The *reductio* argument is confused, however, and we can begin to see why by inquiring about the role the FRM plays. By definition, the FRM is a descriptive property such that whatever has it is right, which is to say is such that, given CTR, the FRM regulates competent use of 'right'. Historically, candidates for the FRM have included such properties as maximizing pleasure or agent-neutral value and compliance with the categorical imperative. One way to unpack the notion that the FRM is "right making" is to say that the FRM just is — or constitutes — the property of being right, and that this constitutive fact explains the identity. On this understanding, both A1 and A2 turn out to be unproblematic. A1 is true because, if the FRM is the property of being right, then the property of being right exists. A2 and the identity together imply that 'right' refers to the FRM.

Long's *reductio* turns on the idea that, if two properties are identical, then it is absurd to think that one property might explain the other. Long (2012: 278) claims that, if the FRM and the property of being right are identical, then "the property that ultimately explains an action's being right [the FRM] just is the property of being right. That is absurd, however: the property that *explains* an action's being right cannot be *identical* to the property of being right" (original emphasis). Long's point might be that if the *explanandum* and the *explanans* are identical, it is absurd to think that we could have an adequate explanation. Since it is the properties and not the explanatory expressions that are supposed to be identical, this charge cannot be quite right. If the explanation we seek is causal, Long might be claiming that CTR and ethical theory together entail that cause and effect are identical. That would indeed be an absurd suggestion. But the explanation ethical theory seeks is the answer to, "what makes actions right", and this question is not about the cause of rightness so

much as its constitution.

Long's argument might be a version of Frege's (1892) puzzle about identity: how can we explain the difference in cognitive significance between $a = a$ and $a = b$? The former seems uninformative compared to the latter. The terms 'morning star' and 'evening star' refer to the same object. But the statement that 'morning star = morning star' is analytic, whereas the statement that 'morning star = evening star' is not. Frege's solution to this problem invokes his famous distinction between sense and reference. The senses of 'morning star' and 'evening star' differ, but they have the same reference. The difference in sense explains the difference in cognitive significance between 'morning star = morning star' and 'morning star = evening star.' The sameness of reference is a consequence of the identity.

Perhaps it is misleading to view Long's argument in light of the identity of 'morning star' and 'evening star', which after all name an object. Moreover, we do not use the morning star as a causal explanation of the evening star, nor would we say that the morning star constitutes the evening star. The FRM and the property of rightness are properties, not objects. The identity of normative and descriptive properties usually receives attention from reductionists, who argue that they are identical because the one is reducible to the other. Schroeder (2005) urges caution in this project: reductionists who assume, for example, that the set of properties to be reduced and the reduction base are complementary and so disjoint appear to contradict themselves. If we define descriptive properties as non-normative properties, then asserting that normative properties are descriptive properties seems to entail a contradiction. Instead, he argues, two modes of reduction seem plausible. The first, and the one I will discuss, is that developed in Jackson 1998, according to which normative properties are reducible to descriptive properties because the former constitute a proper subset of the latter. Since Jackson defines descriptive properties as those that can be picked out by descriptive predicates, his reduction, according to Schroeder (2005: 10), "amounts to the claim that normative properties can be picked out by uncontroversially descriptive predicates. *This* is a perfectly coherent view" (original emphasis).

A view like Jackson's can underwrite an explanatory relation between descriptive properties and the normative properties to which

they are identical. Ethical theorists seek a FRM that is identical to the normative property of rightness. Suppose value-maximizing is the (descriptive) FRM, and suppose that Jackson is right to think that the normative property of rightness is reducible to a descriptive property. It follows that the properties are identical and that rightness is value-maximizing. Moreover, the descriptive predicate 'value-maximizing' picks out the normative property of rightness. Far from being impossible or absurd as Long claims, that result would be informative and illuminating, since it would explain why maximizing value is right. The identity of value-maximizing with rightness accommodates Long's assumptions A1 and A2 because the property of being right exists and the predicate 'right' refers to it. On this view 'right' refers to the property of rightness, which is also the property of maximizing value. The view's explanatory power lies in linking the descriptive predicate to the normative property, not in anything mysterious about the identity. And if Jackson is right, the fact that the descriptive predicate 'value-maximizing' and the normative predicate 'right' both refer to the same property should hardly be surprising: his thesis is that normative properties are a subset of descriptive properties, and thus that *all of them* may be picked out by both normative and descriptive predicates.[3]

The explanation of rightness in terms of the FRM emerges from linking those predicates in certain systematic ways justified by ethical theory. At issue here could be the sense in which the FRM is "right making", where the FRM constitutes rightness. The relation of the FRM to rightness represents a point of contact between ethics and metaethics. Ethics has an interest in the identity between the FRM and rightness in virtue of its need to explain rightness in terms of the FRM. Such an explanation is useful both practically, by pro-

[3] Schroeder's preferred mode of reductionism does not offer a further alternative to thinking that the identity of normative and descriptive properties must be explanatorily inert. He proposes that we could, for example, reduce normative to descriptive properties through analysis rather than, as Jackson does, by regarding one as a subset of the other. Schroeder regards this difference as a strength, since it enables him to define descriptive properties as non-normative and yet reduce the normative to the descriptive without contradiction. As intriguing as it is, his view would not yield property identity, which is the sticking point in Long 2012. For more detail, see Schroeder (2005: 10ff.).

viding guidance to decision making, and epistemically, by offering resources for justifying action. That is the ethical perspective on the question, 'what makes an action right?' Metaethics has an interest in the identity between the FRM and rightness in virtue of its need to explain the semantic value of the predicate 'rightness'. The ethical issue is a question in the metaphysics of morality, since it requires accounting for the sense in which the FRM constitutes rightness. The metaethical issue is a question in the semantics of moral language, since it deploys the identity of the FRM and rightness in order to explain the semantic values of the corresponding predicates.

This point of overlap is important in the context of recent discussions concerning the relation of ethics and metaethics. Indeed, Dworkin (2011) argues vigorously in favor of collapsing the distinction altogether because, as Kalderon (2013) points out, he thinks that all significant metaphysical questions ought properly to be conceived as first-order and substantive. The murky metaethical waters of constructivism are beyond our scope here, but it is an interesting question whether some similar argument shows that all significant semantic questions ought likewise to be conceived as first-order and substantive. That would be the relevant point to establish with relation to the identity of the FRM with rightness. I can only gesture at a negative answer: Putnam's (1976) discussion of the synthetic identity of 'temperature' with 'mean molecular kinetic' energy presupposes a result in physics, but it would be a stretch to conclude that it is therefore a contribution to physics. I suspect that a similar conclusion might be reached with regard to the semantics of rightness given CRT. Though CRT presupposes the identity of the FRM with rightness, the account of the semantic values of the corresponding predicates might not thereby constitute a contribution to ethical theory. At least, we should await an argument that shows why it should do so.[4]

Michael Byron
Philosophy Department
Kent State University
PO BOX 5190

[4] I am grateful to an anonymous reviewer, whose suggestions and comments substantially improved this paper.

Kent OH 44242–0001
USA
mbyron@kent.edu

References

Dworkin Ronald. 2011. *Justice for Hedgehogs*. Cambridge, MA: Harvard University Press.

Frege, Gottlob. 1892.Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100: 25-50. Translated as On Sense and Reference by M. Black in *Translations from the Philosophical Writings of Gottlob Frege*. Edited by P. Geach and M. Black. Oxford: Blackwell, third edition, 1980.

Kalderon, M. A. 2013. Does metaethics rest on a mistake? *Analysis* 73: 129-138.

Jackson, Frank. 1998. *From Metaphysics to Ethics*. Oxford: Clarendon Press.

Long, Justin. 2012. Right-Making and Reference. *American Philosophical Quarterly* 49: 277-280.

Putnam, Hilary. 1979. On Properties. In *Philosophical Papers: Vol. 1, Mathematics, Matter and Method.* Cambridge: Cambridge University Press, 305-322.

Railton, Peter. 1986. Moral Realism. *Philosophical Review* 95: 163-207.

Russell, Bertrand. 1905. On Denoting. *Mind* 14: 479-493.

Schroeder, Mark. 2005. Realism and Reduction: the Quest for Robustness. *Philosophers' Imprint* 5: 1-18.

Shafer-Landau, Russ. 2003. *Moral Realism: A Defence*. Oxford: Oxford University Press.

Streumer, Bart. 2011. Are Normative Properties Descriptive Properties? *Philosophical Studies* 154: 325-348.

Sturgeon, Nicholas. 1988. Moral Explanations. In *Essays on Moral Realism*. Edited by G. Sayre-McCord. Ithaca and London: Cornell University Press, 229-255.

Minimal Semantics and Word Sense Disambiguation

Luca Gasparri
Institut Jean Nicod – ENS Paris

BIBLID [0873-626X (2014) 39; pp. 147-171]

Abstract
Emma Borg has defined semantic minimalism as the thesis that the literal content of well-formed declarative sentences is truth-evaluable, fully determined by their lexico-syntactic features, and recoverable by language users with no need to access non-linguistic information. The task of this article is threefold. First, I shall raise a criticism to Borg's minimalism based on how speakers disambiguate homonymy. Second, I will explore some ways Borg might respond to my argument and maintain that none of them offers a conclusive reply to my case. Third, I shall suggest that in order for Borg's minimalism to best accommodate the problem discussed in this paper, it should allow for semantically incomplete content and be converted into a claim about linguistic competence.

Keywords
Semantic minimalism, lexico-syntactic processing, literal meaning, word sense disambiguation, homonymy.

1 Introduction

Emma Borg (2004, 2012) has characterized semantic minimalism as the natural inheritor of a formal semantics approach to sentential meaning and has defended the idea of a purely lexico-syntactic route to propositional content. In her view, literal content for well-formed declarative sentences is truth-evaluable, fully determined by their lexico-syntactic features, and recoverable by language users with no need to access contextual information or world knowledge. Sentences have their truth-conditional content determined independently of non-linguistic factors, and the contribution of context to the recovery of sentential meaning is limited to the saturation of a

Disputatio, Vol. VI, No. 39, November 2014
Received: 15/12/2013 Revised: 09/02/2014 Accepted: 28/07/2014

narrow class of indexical expressions. Semantic minimalism thus opposes contextualist, relativist and occasion-sensitive views maintaining that the bearers of propositional content are utterances, rejects the all-pervasive constructive role for non-linguistic context envisaged by dual pragmatics, proposes that the proper task of semantic theories is to account for the literal meaning of sentences rather than for the communicated content of speech acts, and is committed to an orthodox view of compositionality, according to which, barring explicit indexicals, the truth-evaluable content of sentential expressions is entirely a function of the combination of their syntactic architecture with the stable semantic input of their lexical constituents.[1]

Understood along these lines, semantic minimalism is vulnerable to two fundamental arguments (Borg 2012: 48-49):

(i) *Minimal propositions are explanatorily inert*: literal truth-evaluable content plays no indispensable role in (accounting for) the cognitive processes whereby speakers assign pragmatic meaning to declarative sentences;

(ii) *Minimal propositions are impossible*: some declarative sentences fail to convey (or encode) literal truth-evaluable content thanks to their lexico-syntactic elements alone.

To start, let us focus on argument (i). The bulk of the objection (a clear formulation of which can be found, e.g., in Recanati 2004: 18-22) is that the entertainment of literal truth-evaluable content needs not be always included in the series of mental processes whereby speakers recover speech act content or pragmatically enriched meaning. If on a hot summer night I tell my thirsty friend John 'There is beer in the fridge', it seems there is no need for him to consciously or unconsciously entertain the literal proposition THERE IS BEER IN THE FRIDGE in order to understand that the beer I am talking about is presumably contained in cans or bottles, rather than spilled everywhere in the fridge. In other words, John needs not entertain overt quantification to determine intended quantification: he can get straight to

[1] For an overview of the main tenets of Borg's minimalism, see Borg 2007, 2009. Korta, Perry 2006, Jaszczolt 2007 and Stojanovic 2008 are equally useful introductions to the positions surrounding the debate on semantic minimalism.

intended quantification, with no intermediate literal stops. Something similar appears to happen in the spontaneous enrichment of sentences with unarticulated content (e.g., 'The baby cried and the mother [+ OF THE BABY] picked it up'), in the recovery of the intended meaning of sentences requiring context-driven quantifier domain restrictions (e.g., 'There is no one at work, everyone went to the party'), in the interpretation of adjectives whose conventional semantic value is *prima facie* unable to function as a propositional constituent prior to contextual input (e.g., 'Paul is ready': for what?), in the comprehension of sentences containing presuppositions that fail to be accommodated by the context of utterance (e.g., 'The dog is thirsty' when no salient dog can be identified),[2] or in the evaluation of sentences with gradable predicates (e.g., 'Mary is tall').[3] Since, the argument goes, these cases prove that the entertainment of literal truth conditions is sometimes unnecessary to determine pragmatically enriched content, semantic minimalism is wrong in requiring the composition of literal truth conditions to take place even in cases where minimal propositions make no contribution to the recovery of pragmatic meaning.

The rationale of this paper is that Borg's minimalism is susceptible to the same variety of objection even if we focus on the determination of literal sentential meaning itself, and that there is no need to

[2] As the informed reader will know, Frege and Strawson proposed that in similar cases the sentence fails to result in a logical form capable of having a truth value. This view of presuppositions is well-established in linguistics: see Heim 1983, Van der Sandt 1992, Beaver 2001.

[3] In this case, the argument from the contextualist side runs as follows. Propositions have truth values relative to circumstances of evaluation. If circumstances of evaluation are possible worlds, then propositions have truth values relative to worlds (i.e., intensions). So if there is a minimal proposition literally expressed by every utterance of 'Mary is tall' at every context of use, 'Mary is tall' must have an intension. At this point, contextualists conclude *modo tollente* that since the gradable adjective *tall* makes it impossible for 'Mary is tall' to have an intension in the standard sense of the term, there is no stable proposition literally expressed at every utterance of 'Mary is tall'. For an attempt to address the problem in a synthesis of semantic minimalism and radical contextualism, see the non-indexical contextualism of MacFarlane 2007, 2009. See Davis 2013 for a fresh discussion of MacFarlane's proposal. More on the semantics of gradable adjectives in, e.g., Kennedy 2007.

point at the interplay of literal truth conditions and pragmatically enriched meaning to maintain that minimalism fails to match our best assumptions about the role of truth-evaluable content in meaning recovery. Bearing this in mind, my type (i) aim will be to argue that there are well-formed declarative sentences such that they cannot be assigned a literal truth-evaluable content through lexico-syntactic processing alone. This will lead me to the contention that there are cases of sentential meaning construction where the recovery of truth-evaluable content cannot be based exclusively on linguistic knowledge and, in turn, to the type (ii) claim that sentences whose literal meaning can be determined only via an interactive procedure of the above kind fail to convey a literal truth-evaluable content unless they are processed on the basis of information exceeding the boundaries of linguistic knowledge (§2). I shall explore some ways the minimalist might respond to this objection and argue that none of them offers a conclusive reply to my case (§§3,4). Finally, I will suggest that Borg's minimalism should allow for semantically incomplete content and be converted from a thesis about lexico-syntactic performance into a claim about lexico-syntactic competence (§5).

2 Homonymy

Consider the sentence 'A pupil was in the middle of the classroom'. Due to the presence of *classroom*, speakers exposed to this sentence preferentially interpret the homonymous word *pupil* as conveying the sense YOUNG STUDENT, despite the fact that *pupil* can also be taken to mean EYE OPENING. Accordingly, they tend to perceive 'A pupil was in the middle of the classroom' as a semantically definite expression despite the ambiguity of *pupil*. They might perceive *pupil* as ambiguous in the earliest stages of the speech input, when they have not yet been provided with any clue as to how *pupil* should be disambiguated. But as soon as they get to hear *classroom*, they spontaneously select YOUNG STUDENT as the most plausible sense to be assigned to *pupil*. The whole process runs plausibly (and very roughly) as follows.[4] First,

[4] For the relevant empirical research, see any recent handbook of psycholinguistics with a good section on lexical processing (e.g., Traxler 2011, Spivey, McRae, Joanisse 2012 or Harley 2014). Klepousniotou 2002 and Morris 2006

the content words and the functional items occurring in 'A pupil was in the middle of the classroom' are linearly paired with a provisional semantic interpretation. Unambiguous words are paired with a single sense, whereas *pupil* is paired with both its alternative senses, and all the senses thus activated are deposited in working memory. Parallel morphosyntactic processing supervises the construction of the phrase structure for the activated senses and yields the proto-proposition A [YOUNG STUDENT / EYE OPENING] WAS IN THE MIDDLE OF THE CLASSROOM. In order to associate 'A pupil was in the middle of the classroom' with a definite set of truth conditions, the speaker now needs to select one of the two candidate interpretations of *pupil*. To this end, A [YOUNG STUDENT / EYE OPENING] WAS IN THE MIDDLE OF THE CLASSROOM is transferred into a post-semantic workspace which contrasts the statistical likelihood of A YOUNG STUDENT WAS IN THE MIDDLE OF THE CLASSROOM with that of AN EYE OPENING WAS IN THE MIDDLE OF THE CLASSROOM, pares away EYE OPENING as otiose, selects YOUNG STUDENT and delivers the truth-evaluable string A YOUNG STUDENT WAS IN THE MIDDLE OF THE CLASSROOM.

The exact nature of the selection occurring at the end of this process is not immediately relevant for our purposes: it could be a statistical inference computing on the frequency of the senses assigned to *pupil* in previous occasions of use, or it could involve the access to some rule-like constraint binding the interpretation of *pupil* to YOUNG STUDENT whenever *pupil* is used in a sentence containing a relevant occurrence of *classroom*. Regardless of this, there is a single important point to be made for the purposes of our argument. Namely, post-semantic selection is indispensable to pair the input sentence with a definite truth-evaluable content: without the reduction of [YOUNG STUDENT / EYE OPENING] to YOUNG STUDENT, the comprehender is bound to be unable to associate 'A pupil was in the middle of the classroom' with a single set of truth conditions. Albeit I doubt that the reader has ever stepped into a classroom and spotted a bare human eye right at the center of it, there is plenty of conceivable contexts of utterance where construing *pupil* as EYE OPENING

---

both offer a nice introduction to the psychology of word sense disambiguation. Small, Cottrell, Tanenhaus 1988 is also a comprehensive, though earlier reference work on the subject.

would make it an impeccable referential label for an object standing in the middle of a classroom (similar cases are discussed, e.g., in Searle 1980, Pelczar 2000, Recanati 2004). Simply put, there are no linguistic reasons why the interpreter should prefer YOUNG STUDENT over EYE OPENING, and it is impossible to require that the information whereby agents select YOUNG STUDENT be part of their command of the lexico-syntactic properties of English. The discriminating factor, here, is world knowledge.[5] Contra Borg, there seems to be no definite "what is said"[6] without appealing to information outside the language faculty here, because unless the two candidates to the status of truth-evaluable content conveyed by 'A pupil was in the middle of the classroom' are tested against a background of relevant non-linguistic knowledge, it is impossible to assign the sentence a single literal meaning.[7]

Now, while many accept that encoded conventional meaning is typically non-propositional and observe that in most cases disambiguation and reference determination are needed to obtain truth conditions (e.g., see Devitt 2013), semantic minimalism wants encoded

[5] A quick counterexample might be useful to clarify this point. The verb *find* is highly polysemous: it can express LOCATE, BELIEVE, REALIZE and plenty of other fine-grained senses. Suppose we need to interpret the sentence 'Mark found that the show was boring' and are asked to choose which, among LOCATE and BELIEVE, is the sense to be assigned to *find*. To do this, we do not need world knowledge, because it is part of our word-level command of the combinatorial properties of *find* that when the object slot of its argument structure is filled by a sentential complement, the verb cannot be interpreted as LOCATE (as in, e.g., 'Mark found the cat'). The situation is different in 'A pupil was in the middle of the classroom': in this case, a non-linguistic input is indispensable to perform sense selection.

[6] By 'what is said', I simply mean the conventional truth-conditional features that can be ascribed by a speaker A to a sentence S in virtue of the linguistic properties of S (hence, in virtue of A's being a competent user of the language in which S is expressed).

[7] Naturally, the example I have chosen is just one among many possible instances of homonymy, both balanced (i.e., based on word forms licensing equally dominant senses, such as *cell* or *panel*) and unbalanced (i.e., based on word forms whose alternative senses are asymmetric in frequency, such as *ball* or *port*). More precisely: in 'A pupil was in the middle of the classroom', *pupil* is a balanced homonym occurring in a biased sentential context, that boosts the statistical likelihood of YOUNG STUDENT.

conventional meaning to be inherently propositional. Which makes disambiguation a problematic case. Borg (2004: 140-146; 2012: 90-91, 171-172) lucidly recognizes the issue and provides some nicely argued answers regarding how it should be accommodated in the context of her minimalist proposal. In what follows, I will argue that none of such answers is entirely convincing. To be fair, I will never claim to have identified a knock-out case against Borg's thesis, but I think I can reasonably show that the best assumptions we can make about the dynamics of word sense disambiguation cast some significant doubts on the overall plausibility of minimal semantics. To proceed, let us examine how Borg suggests that her theory can accommodate cases of lexical ambiguity of the sort contemplated in 'A pupil was in the middle of the classroom'. According to Borg, minimal semantics can deal with them because disambiguation processes typically fall into one of the following cases.

(D1) *Pre-Linguistic Disambiguation*. Sense selection occurs *before* lexico-syntactic processing. Only one of the two senses of *pupil* is inputted to lexico-syntactic processing and only one of the two truth-evaluable contents licensed by the sentence is built.

(D2) *Post-Linguistic Disambiguation*. Sense selection occurs *after* lexico-syntactic processing. The sentence is heard as ambiguous and both its alternative truth-evaluable contents are built. After the two truth-evaluable contents have been allowed to leave the language faculty, general intelligence selects one and suppresses the other.

(D3) *Linguistic Disambiguation*. Sense selection occurs *inside* lexico-syntactic processing. This can happen in three ways.

(D3a) Both senses of the homonym are inputted to lexico-syntactic processing but only one is used to interpret the sentence, due to a habitualized preference. For example, the subject's previous encounters with the homonym have established a selectional tendency based on which her language faculty spontaneously computes one of the two senses and pares away the other.

(D3b) Both senses of the homonym are inputted to lexico-syntactic processing but one is immediately suppressed thanks to knowledge about the preceding discourse context.

(D3c) Both senses of the homonym are inputted to lexico-syntactic processing and used to interpret the sentence. General intelligence intrudes into lexico-syntactic processing and operates as a selective inhibitor on one of the two truth-evaluable contents licensed by the sentence, before they are allowed to leave the language faculty.

3.1 Pre-linguistic disambiguation

Let us start with (D1): only one of the two senses of *pupil* is inputted to lexico-syntactic processing. Borg's proposal can be spelled out in two ways. First, preferential sense activation is direct, unconstrained and obtains independently of the general intelligence of the interpreter (let us name this hypothesis D1a). Second, preferential sense activation is the output of the early resolution of a constraint-satisfaction problem and obtains thanks to the general intelligence of the interpreter (let us name this hypothesis D1b). Opting for (D1a) would clash with a good deal of classic experimental literature on lexical access observing that in the earliest stages of language interpretations tasks, all the alternative senses of an ambiguous word are activated (regardless of which is more dominant or contextually appropriate) to be later selected via the acquaintance with semantic and non-linguistic context (e.g., Swinney 1979; Tanenhaus, Leiman, Seidenberg 1979; Seidenberg et al. 1982; Folk, Morris 2003; Mason, Just 2007). As for (D1b), the hypothesis could be reconciled with Borg's minimalism on condition that the general-purpose processes involved in pre-semantic filtering required no input from lexico-syntactic analysis, i.e., that pre-semantic filtering and lexico-syntactic analysis were two serially ordered processes among which there occurred a rigidly unidirectional interaction. Yet, it seems that in order for constrained sense activation to obtain, a rather rich flow of information from pre-semantic filtering to lexico-syntactic analysis and back from lexico-syntactic analysis to pre-semantic filtering has

to take place. To appreciate this, let us observe that there are two essential ways in which Borg's appeal to (D1b) can be put to work. The first would be to assume that constrained sense activation is generated by a world knowledge default (§3.1.1). The second would be to opt for a salience-first model of lexical access (§3.1.2).

3.1.1 World knowledge defaults

Assume that the retrieval of the senses of *pupil* is controlled by a world knowledge default (WKD) that constrains lexical access to the activation of YOUNG STUDENT by computing on information provided in the position occupied by *classroom* (easy test: were the source sentence 'A pupil was in the middle of an eye', the interpretation accessed for *pupil* would have been EYE OPENING).[8] Now ask: is it possible for speakers to recruit the WKD and bring it to bear on the operations of the language faculty before lexico-syntactic processing has started? The answer to this question seems bound to be negative, since without some early lexico-syntactic breakdown of the sentence it is impossible for speakers to determine that the syntagmatic position occupied by *classroom* is the one containing the information that is relevant for the disambiguation of *pupil*. If it is true that subjects selectively retrieve YOUNG STUDENT because of a WKD based on a stereotypical representation of the objects they are more likely to find in classrooms, it is also true that the WKD can be brought to bear on the disambiguation of *pupil* only if some rudimentary construc-

[8] Incidentally, let it be noted that the default can be modeled with the toolkit of standard information theory. The idea is to measure the amount of information required to disambiguate among candidate interpretations for an ambiguous word in terms of entropy generated by candidate interpretations in the set of its senses, and posit that only the interpretation generating the lowest entropy will be activated. If H is the entropy, P is a measure of probability, M is the set of the senses of *pupil*, C is the set of contextual bits of information available in the interpretation of 'A pupil was in the middle of the classroom', *cl* is *classroom*, YS is YOUNG STUDENT, and EO is EYE OPENING, then the preferential activation of YOUNG STUDENT can be predicted on grounds that $H[YS|C] < H[EO|C]$, which is to say: $-\sum_{cl\in C} P(cl) \sum_{YS\in M} P(YS|cl) \log P(YS|cl) < -\sum_{cl\in C} P(cl) \sum_{EO\in M} P(EO|cl) \log P(EO|cl)$. More generally, *classroom* is a relevant bit of contextual information because $cl\in C$ and $H[M] > H[M|C]$, that is, $-\sum_{m\in M} P(m) \log P(m) > -\sum_{cl\in C} P(cl) \sum_{m\in M} P(m|cl) \log P(m|cl)$, with *m* being an interpretation of *pupil*. See Cover, Thomas 2006.

tion of the logical form of the sentence and some exploratory analysis of its lexical items have already established at least what follows: (i) that *pupil* is an ambiguous item describing the value of the predicate variable of an existential quantifier; (ii) that *classroom* is the most salient element of an optional argument signaling the location of the object labeled as *pupil*; (iii) that *classroom* "means" CLASSROOM. Unless this body of information has been processed by the lexico-syntactic parser, general intelligence is bound to be unable to determine which among the many WKDs the agent is acquainted with should be accessed and exploited for preferential sense activation. In this sense, (D1b) can be treated as a variant of (D3c), and its viability comes to depend on the possibility of reconciling Borg's minimalism with the existence of a dense informational interface between lexico-syntactic processing and general intelligence.

3.1.2 Salience-first access

Here the minimalist might want to observe that the psychological story we have recapitulated in §2 is not completely accurate or unbiased, since some evidence has been taken to shown that in the earliest stages of utterance processing ambiguous words do not always activate the entire range of their alternative senses, but are sometimes paired with a single "contextually salient" meaning (see, e.g., Giora 2003, 2012). If the rest of the phrase structure accepts the salient meaning (i.e., the incorporation of the salient meaning does not generate syntactic or semantic anomalies), the salient meaning is preserved and the string is allowed to leave the language faculty. Otherwise, if the rest of the phrase structure does not support the salient meaning (e.g., the salient meaning makes a predicate invalidate a selectional restriction or pushes a complement to violate the argument structure of its verb), the lexico-syntactic module activates a backtracking function which inhibits the salient meaning and recruits a more appropriate sense for the ambiguous word. Now, if we claim that the preferential activation of the salient sense is stimulated directly by the context of utterance, we face the same problem encountered in the discussion of (D1a) (i.e., in the earliest stages of language interpretations tasks, all the alternative senses of an ambiguous word are activated regardless of which is contextually appropriate).

On the other hand, if we construe salience as a feature engendering preferential sense activation on the basis of the interpreter's previous linguistic experience (i.e., all the alternative senses of the ambiguous word are activated but only the most prominent has enough statistical strength to get to the composition phase), we face two different problems. First, by allowing that linguistic experience can have a constructive role in arranging the dominance of the senses associated to the mental representations of word forms, Borg would endorse a notion of lexical knowledge which is much more idiosyncratic and open-ended than the formalist picture of linguistic competence she aims to safeguard (2012: chapter 6). Second, it appears that the string built by lexico-syntactic processing with the dominant sense of the homonym must be verified by an additional abductive stage in order to be definitively validated. The most indicative reason for this is that backtracking can be requested even after the string built by the lexico-syntactic parser has left the linguistic module. Suppose we are dealing with a native speaker of English S who is informed about the ambiguity of *pupil* and whose previous linguistic experience is such that S has been prevalently exposed to uses of *pupil* where the word was intended to mean EYE OPENING. At some point, S is presented with 'A pupil was in the middle of the classroom'. S detects the word *pupil*, preferentially accesses EYE OPENING and integrates it in the phrase structure of the sentence. Is this sufficient to predict that S will validate the string constructed by interpreting *pupil* as EYE OPENING? Of course not. S could still evaluate that the string including EYE OPENING embeds a heuristically questionable interpretation of the sentence because the reading she has spontaneously assigned to *pupil* displays an insufficient degree of encyclopedic consistency with the rest of the sentential context, and activate backtracking to replace it with a less atypical interpretation. The moral is simple: when it comes to cases where the competitor interpretations for an ambiguous word generate truth-conditional strings which are equally acceptable if tested against the *desiderata* of lexico-syntactic well-formedness, it is impossible to require the language faculty to be responsible for the unreflective sense of implausibility a speaker accessing EYE OPENING first could feel in evaluating whether the interpretation she has assigned to our sentence is attractive. Yet, it seems that such a stage of weighted abduction is a constitutive part of the array of mental

processes whereby cognizers assign literal content to sentences, and it is not entirely clear how this could be reconciled, to borrow Borg's (2004: 142) own words, with the somewhat "ascetic" postulation of a self-contained lexico-syntactic route to sentential meaning.

3.2 Post-linguistic disambiguation

Let us now turn to (D2). Borg depicts the following scenario: the sentence 'A pupil was in the middle of the classroom' is heard as ambiguous and both its alternative truth-evaluable meanings are built. After the two truth-evaluable strings have been allowed to leave the language faculty, general intelligence selects one and suppresses the other. Now, I have no complaints against the idea that disambiguation might sometimes follow this routine, but I do have some doubts regarding the compatibility between the existence of processes of post-linguistic disambiguation and Borg's minimalism. It is true that in cases of this sort lexico-syntactic calculus does a fundamental part of the job required to construct the truth-evaluable content speakers preferentially associate with the target sentence. Be it made of a single structured string with an underspecified lexical slot (A [YOUNG STUDENT / EYE OPENING] WAS IN THE MIDDLE OF THE CLASSROOM), be it made of two fully truth-evaluable strings to be evaluated and selected as wholes (A YOUNG STUDENT WAS IN THE MIDDLE OF THE CLASSROOM VS. AN EYE OPENING WAS IN THE MIDDLE OF THE CLASSROOM), the acquaintance with the lexico-syntactic features of our sentence yields some semantically informative content. Yet, by dropping the tenet that the truth-evaluable content of well-formed declarative sentences is always and entirely dictated by lexico-syntactic processing, minimalism would dispose of the key claim thanks to which it promised to offer an interesting and controversial insight on language processing in the first place. The point of Borg's minimalism is not to merely suggest that lexico-syntactic calculus and linguistic knowledge have a crucial and modularly defined role in the recovery of sentential meaning, or that they are "usually" sufficient to obtain truth-evaluable content. The point of Borg's minimalism (e.g., 2012: 48) is to propose that lexico-syntactic calculus and linguistic competence are *all it takes* to build semantically complete content, and it is difficult not to remain skeptical about how the case we have discussed is sup-

posed to meet this Davidsonian *desideratum*. The very choice of labeling comparable processes of disambiguation as 'post-semantic' looks unwarranted. Granted, they are supposed to intervene on the output of the Fodorian module Borg assigns to lexico-syntactic processing and, in this sense, their temporal niche naturally falls in the post-modular inferential phase. But if it is true that they are indispensable to generate (rather than modulate or enrich) the literal content of the sentence in need of disambiguation, is it still fair to characterize them as something *bona fide* "post-semantic"?

3.3 Linguistic disambiguation

Finally, let us turn to (D3). Borg suggests that it is possible to locate disambiguation within lexico-syntactic processing while preserving the minimalist framework in three scenarios, respectively corresponding to (D3a), (D3b) and (D3c). In the first case, both the senses of *pupil* are inputted to lexico-syntactic processing but only one is used to interpret the sentence, due to a habitualized preference. In the second case, both the senses of *pupil* are inputted to lexico-syntactic processing, but one is rapidly suppressed thanks to information acquired from the preceding discourse context. In the third case, both the senses of *pupil* are inputted to lexico-syntactic processing and used to interpret the sentence, but general intelligence acts as a selective inhibitor on one of the two truth-evaluable contents licensed by the sentence while they are being built.

As for (D3a), it seems that adhering to this line of reply would again relax the notion of 'linguistic knowledge' beyond the limits tolerated by Borg's commitment to a formalist understanding of natural language semantics. As we observed in §3.1.2, by accepting that factors such as conventionality, distributional frequency and familiarity can affect the offline dominance of the senses associated to word forms in the mental lexicon, the minimalist would endorse a view of lexical competence which is much more "pragmatic" than the one its intellectualist understanding of semantic competence should aim to sustain. More precisely, if the representational repertoire underpinning our ability to make competent use of word forms contained instructions of type '(when embedded in the semantic context C) the word *w* preferentially takes the sense *m*', word knowledge would

more closely mirror the abductively rich notion of lexical semantic knowledge adopted by cognitive linguists (e.g., Evans 2009) and by most NLP approaches to word sense disambiguation (e.g., Manning, Schütze 1999; Agirre, Edmonds 2006), rather than the axiomatic-logicistic picture of lexical meaning envisaged by Borg (e.g., 2010).

As for (D3b) and (D3c), I suspect that allowing linguistic performance to be inhibited by the discourse context would jeopardize Borg's appeal to a modularist account of lexico-syntactic processing. Borg (2004: 93, 144; 2012: 64) is right in observing that allowing information about the discourse context to play a role as a selective top-down filter on meaning construction processes is very different from making room for the kind of all-pervasive constructive role for non-linguistic context envisaged by some contextualists. Even so, the "intrusion" licensed by Borg seems far from innocent here: however weakly one chooses to interpret it, it remains based on the assumption that the core computations of the lexico-syntactic module can be stopped by the short-term representations in which the interpreter has stored the background information she has acquired from the discourse context. This would be absolutely fine in a pragmatically-oriented approach such as discourse representation theory (Kamp, Reyle 1993), segmented discourse representation theory (Asher, Lascarides 2003), or file change semantics (Heim 1988), but seems to clash with Borg's appeal to a full-fledged Fodorian picture of lexico-syntactic processing (which includes the *mandatoriness* requirement: in order for some system to constitute a module, its operations must run to completion every time they are switched on by presentation of a relevant stimulus; see Fodor 1983).[9] Hence, even if the susceptibility to inhibition displayed by linguistic performance

[9] Other studies support the same worry. For example, Bicknell et al. 2010 and Matsuki et al. 2011 report an increase in reading times for sentences in which an agent-verb combination is followed by a statistically incongruent (though linguistically plausible) patient (e.g., 'The journalist checked the spelling of his latest report' vs. 'The mechanic checked the spelling of his latest report'). The immediacy of this slowdown would seem to require either that world knowledge must be embedded in the lexicon, or else that world knowledge can affect the amount of time required to carry out the analysis of the linguistic properties of a sentence by manipulating the operations of the lexico-syntactic parser while they are being performed. Both options look problematic for minimal semantics.

in word sense disambiguation were as selective as Borg believes it is, this could still be sufficient to conclude, contra the minimalist, that lexico-syntactic processing should be better characterized as a weakly modular system, i.e., domain-specific and designed to contribute to the construction of truth-evaluable content by interfacing its operations with information exceeding the knowledge base of the lexico-syntactic parser. Ironically enough, the very idea Borg is so concerned with rejecting (i.e., broadening the data base of linguistic processing so as to make it compute information outside the lexico-syntactic province) might be the best way to safeguard the minimalist thesis that lexico-syntactic processing is systematically sufficient to assign sentences a truth-evaluable content via a fully modular information processing routine.

4 Multiplying types

I add to the list of possible responses an answer that Borg does not consider explicitly, but that I find worth articulating and discussing. Let me express it in the form of a question: why not conceive A YOUNG STUDENT WAS IN THE MIDDLE OF THE CLASSROOM and AN EYE OPENING WAS IN THE MIDDLE OF THE CLASSROOM as the (minimal) propositions expressed by two different sentence-types whose respective English instantiations happen to be phonographically indiscernible? After all, homonymy is standardly understood as an *n*-ary relation between different terms that share the same spelling and pronunciation (e.g., Murphy 2010). So it should stand to reason to argue that the possibility to pair our sentence with two truth-evaluable contents arises because 'The pupil was in the middle of the classroom' is the realization of two sentence-types that in English happen to be expressed through utterances and inscriptions sharing the same surface properties. The argument could run as follows: (i) the distinction between A YOUNG STUDENT WAS IN THE MIDDLE OF THE CLASSROOM and AN EYE OPENING WAS IN THE MIDDLE OF THE CLASSROOM corresponds to the distinction between two sentence-types, α and β; (ii) the logical forms of α and β terminate in the subject position with different NPs (α hosts the constituent $pupil_{\alpha}$, while β hosts the constituent $pupil_{\beta}$); (iii) since $pupil_{\alpha}$ and $pupil_{\beta}$ are homonyms in English, the difference between the sets of lexical types respectively hosted by α and β can-

not be expressed in the surface realization of α and β; (iv) based on (i-iii), disambiguating *pupil* actually amounts to determining whether we should process α or β; (v) once the relevant sentence-type (α or β) has been selected, its logical form can be recovered via the information processing routine evoked by the minimalist. The argument looks viable and the minimalist can endorse it, perhaps reinforcing it via psycholinguistic claims about homonyms' having a separate mental representation for each of their alternative senses (see, e.g., Frazier, Rayner 1990; Klein, Murphy 2001; Beretta, Fiorentino, Poeppel 2005; Brown 2008), and via a suitably strengthened version of the hidden homonymy approach to color terms ambiguity argued by Kennedy and McNally (2010), a line of argument Borg herself (2012: 91) is sympathetic with. The problem is that since the English instantiations of α and β are indiscernible and therefore bound to be perceived as a single phonographic input, the theoretical distinction between the two sentence-types cannot be used to increase the plausibility of Borg's account of the psychology of language processing. As the disambiguation between α and β cannot be performed on the basis of phonological, presuppositional, graphic or intonational variables, *ceteris paribus* their presentation is bound to be perceived as the presentation of the same sentence and to trigger the same cognitive responses across occasions of interpretation.

To have a clearer grasp of the problem, think of quantifier raising in structurally ambiguous sentences such as 'Some boy loves every girl' (let us name this sentence S). Here we have one surface form admitting two logical forms (LFs), depending on which of the two quantifiers is assigned the wider scope ($\exists > \forall$ vs. $\forall > \exists$). In this case, the argument from the minimalist side could run as follows. Since every well-formed sentence-type is paired with just one LF, the possibility to dislocate the quantifiers of S from their surface position to their scope position in two ways proves that S is the surface realization of two sentence-types. Accordingly, the assumption of a purely lexico-syntactic route to sentential meaning can be preserved on grounds that it cannot be the case that the cognitive processes involved in the treatment of the two types are the same: they are distinct and each consistent with the minimalist view of sentence pro-

cessing.[10] Once more, the reply looks viable, but it seems it underestimates an important detail: unless some independent input as to which of the two quantifier dislocations should be favored is given, it is impossible to determine the literal content conveyed by S (*qua* phonographic input) by relying only on the lexical and syntactic features that are manifested at the surface level. It is more than reasonable to expect the cognitive processes involved in the construction of its two interpretations to be different, but neither of the two can be preferentially triggered if context does not provide some disambiguation clue signaling which of the two types fits the conversational setting. Which means, in turn, that the minimalist can count on this line of reply on condition that lexico-syntactic processing operates directly on disambiguated sentence-types rather than on surface forms, thereby evading the question of how (if not via context-sensitive lexico-syntactic processing) we identify sentence-types and evaluate their contextual plausibility in the first place. Borg might wish to insist that the case is easy to accommodate within her proposal, since she agrees that non-linguistic information can play a role in letting context help select one of the two sentence-types via the selection of its LF. If this were the case, Borg would still fall short of her aspiration to offer a realistic explanation of meaning recovery: in order to get to truth-evaluable content, you must pass through LF selection, and lexico-syntactic processing alone will not do it. In other words, if semantic minimalism wants to live up to its psychological ambitions and give us a plausible story about how agents competent in a linguistic idiom come to grasp the meaning of sentences expressed in that idiom, it should tell us something about how we go from being exposed to unprocessed collections of phonographic events to the entertainment of truth-evaluable thoughts. Unfortunately, it is difficult to tell how narrowing the framework down to an account of the processes governing the assignation of truth conditions to abstract LFs could produce a cognitively instructive account of semantic performance (or even of a stage of semantic performance).[11]

[10] I am grateful to Emma Borg for pressing me to address this point.

[11] In addition, it should perhaps be observed that while the low semantic overlap characterizing the alternative meanings of a homonym is consistent with an appeal to the distinction between lexical types, the same strategy is unlikely to

Very crudely put, it seems that the minimalist has to choose sides. Either she restricts herself to arguing that her claims are intended to spell out the deductive aspects of the semantics of sentence-types and that her emphasis on 'meaning recovery' should be interpreted as a philosophically informed rationalization of the analytical routine whereby orthodox semanticists look into the truth-evaluable content of logically transparent types, or she bets on the hypothesis that her framework can be exploited to describe the mental processes whereby ordinary speakers assign literal content to sentences in concrete and situated events of language use. If the latter option is the case, then my (admittedly epistemological) worry is that if minimalism does not aim to contemplate in some interesting fashion all that is involved in the extraction of semantic information from surface forms, its appeal to a "purely lexico-syntactic route" to truth-evaluable content runs the risk of insisting on a notion of linguistic processing which has nothing to do with the way we use natural languages.

Part of this plea for empirical responsibility is insightfully recommended by Borg, who adds that although minimalism is open to be disconfirmed by the psychological evidence, such evidence must be about speakers' knowledge of their language, not about the psychological processes whereby speakers assign content to sentences (Borg 2012: 64). Yet, even this *caveat* looks problematic if measured against the rest of Borg's claims. First, it is generally well-established that linguistic knowledge and the dynamics of utterance processing cannot be kept too separate, since any plausible theory of the organization of linguistic knowledge is bound to imply rather specific claims about the way such knowledge operates in language comprehension and production, and any plausible theory of the cognitive processes underlying language use is bound to imply rather specific claims about the organization of linguistic knowledge.[12] But more importantly, if

result pursuable in cases of ambiguity generated by word forms whose possible senses are significantly more related. For example, it would be more controversial to postulate a separate lexical type for each of the different senses of a polysemous verb like *take*. Yet, it seems to me that many of the objections this paper has raised against minimal semantics by focusing on the disambiguation of homonymy could be formulated just as fittingly by considering the processing of polysemy. For issues of space, I cannot elaborate further on this point.

[12] For example, the psycholinguistic research on the mental lexicon is stan-

the only type of evidence that can disconfirm Borg's minimalism is about linguistic knowledge (which is by definition a matter of linguistic competence), does this mean that Borg's proposal should be evaluated exclusively with respect to its ability to account for linguistic competence? In such a case, even if we agreed that minimalism is the best deal on the table to make sense of speakers' command of their language, it is not immediately clear why the endorsement of a minimalist view of lexico-syntactic competence should *ipso facto* be interested in trading in the possibility to pair well-formed declarative sentences with truth-evaluable content just through lexicon and syntax. That we should have a minimalist view of lexico-syntactic knowledge does not necessarily mean that we should have a minimalist theory of lexico-syntactic performance: the two are *very* different animals. The same problem seems to undermine Borg's vindication of the minimalist thesis as the natural inheritor of a modularist view of lexico-syntactic processing (e.g., 2012: 13). That we should have a modularist picture of lexico-syntactic processing does not necessarily mean that we should have a modularist view of the entire range of processes involved in the recovery of propositional meaning: lexico-syntactic processing might well satisfy the strong modularity constraint and yet be insufficient to generate truth-evaluable content. For example, one could easily endorse a Fodorian picture of lexico-syntactic processing while maintaining that all the language faculty can do with a well-formed declarative sentence S is to pair S with an abstract template that relates to the truth-evaluable content eventually selected for S just as character relates to content in the variety of contextualism popularized by Kaplan.

5 Concluding remarks

To sum up. I have reviewed the main tenets on Borg's framework and clarified the way the disambiguation of homonymy poses a challenge to the account of language processing proposed by semantic minimalism. I have then examined the solutions offered by Borg in order to accommodate the problem, and argued that none of them

---

dardly understood as an attempt to infer claims about the nature of lexical representations from the study of lexical activity. See, e.g., Jarema, Libben 2007.

offers a fully convincing argument to embrace the austere picture of meaning recovery recommended by semantic minimalism against accounts of linguistic processing where the composition of truth-conditional meaning is vastly context-dependent and richly interfaced with world knowledge and general purpose abilities.[13] Now let me conclude with a couple more constructive proposals. First, I suggest that Borg should opt for a more nuanced picture of the limitations of lexico-syntactic processing and follow the recommendation already argued, among others, by Bach (2007): do without propositionalism. Which means, give up on the claim that every well-formed (indexical-free) declarative sentence expresses a truth-evaluable content which is fully determined by its lexico-syntactic features and recoverable with no need to access non-linguistic information, and make room for the notion of a 'propositional radical' (or for any other construct doing the same explanatory work carried out by propositional radicals, e.g., the schematic truth-conditional templates of relevance theorists). In Borg's (2012: 208) terms: make room for the talk of incomplete logical forms and move toward the "perilous fine line" separating radical minimalism from contextualism. Hence, allow that in some cases the best the acquaintance with the overt lexico-syntactic properties of a sentence can do is pair it with a truth-conditionally incomplete content, while denying that filling the gap required to turn it into a genuine truth-evaluable content has to be based on linguistic knowledge or to be classified as a semantic matter. Second, Borg's minimalism may be much better off as an account of linguistic competence. To my understanding, it is not only possible, but perhaps even desirable to pair a modularly inspired understanding of lexico-syntactic knowledge with the view that the recovery of truth-evaluable content is a phenomenon which is designed to arise via the cooperation of linguistic and non-linguistic information. One simple way to do this could be to introduce a

[13] Although in this article I have restricted myself to questioning Borg's minimalism, the points I have tried to make should apply equally well to Cappelen and Lepore's (2005) minimalism. The reason why I have chosen to confine my discussion to Borg's work is simple. Semantic minimalism owes us an account of its empirical plausibility, but while Cappelen and Lepore tend to shy away from that task, Borg is eager to explore the psychological tenability of her assumptions. Which makes her work a more suitable contact point for my arguments.

distinction between 'knowledge modules' and 'processing modules' (Coltheart 1999), and bet on the research hypothesis that lexico-syntactic analysis is based on weakly modular (i.e., domain-specific and interactive) processor which computes on information retrieved from a strongly modular knowledge base, corresponding to what is standardly referred to as speakers' "knowledge of the language".

After all, if there is one lesson we can plausibly learn from the very existence of ambiguity in natural languages, that lesson is precisely that language processing is designed to interact with general purpose abilities. Functionalist accounts of ambiguity have long been arguing that the presence of ambiguity in natural languages should prove that the language faculty has not evolved for purposes of communication, since, if that were the case, linguistic forms would map bijectively to meanings, and comprehenders would not need to expend effort in inferring literal meaning or speech act content via non-linguistic information (e.g., Pinker, Bloom 1990; Chomsky 2002). On the contrary, it can be argued that ambiguity is in fact a desirable property of communication codes, because it allows for linguistic systems which are formally parsimonious, optimized for memorization and exploitable by drawing computational resources from different domains of cognition. As Piantadosi, Tily and Gibson (2012) have noted, there are two important facts that make ambiguity a desirable feature of linguistic systems and support the intuition that ambiguity results from a pressure for efficient communication. First, when sentential or extra-linguistic context is informative about meaning, a completely unambiguous language would become partly redundant with context and therefore inefficient (no wonder that homonyms and highly polysemous words preserve their ambiguity in a strikingly small proportion of the cases in which they are embedded in a sentence or used in a speech act). Second, by mapping phonographic units to multiple meanings, ambiguity reduces the number and the length of word forms that have to be stored in the mental lexicon to secure efficient linguistic performance, maintains our internal word store at a low degree of phonotactical complexity, and allows language users to speed up the transmission of information via speech despite the physiological limitations of their articulatory apparatus (e.g., Levinson 2000).[14]

[14] Many thanks to Emma Borg, François Recanati, Diego Marconi and Ja-

Luca Gasparri
Institut Jean Nicod, ENS Paris
UMR 8129, Pavillon Jardin
29 rue d'Ulm, 75005 Paris, France
luca.gasparri@ens.fr

References

Agirre, Eneko, and Edmonds, Philip. Eds. 2006. *Word Sense Disambiguation: Algorithms and Applications.* Berlin: Springer.

Asher, Nicholas and Lascarides, Alex. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.

Bach, Kent. 2007. The Excluded Middle. Semantic Minimalism Without Minimal Propositions. *Philosophy and Phenomenological Research* 53: 435-442.

Beaver, David. 2001. *Presupposition and Assertion in Dynamic Semantics*. Stanford, CA: CSLI Publications.

Beretta, Alan, Fiorentino, Robert and Poeppel, David. 2005. The Effects of Homonymy and Polysemy on Lexical Access: An MEG Study. *Cognitive Brain Research* 24: 57-65.

Bicknell, Klinton, Elman, Jeffrey L., Hare, Mary, McRae, Ken and Kutas, Marta. 2010. Effects of Event Knowledge in Processing Verbal Arguments. *Journal of Memory & Language* 63: 489-505.

Borg, Emma. 2004. *Minimal Semantics*. Oxford: Oxford University Press.

Borg, Emma. 2007. Minimalism versus Contextualism in Semantics. In *Context-Sensitivity and Semantic Minimalism: New Essays on Semantics and Pragmatics*. Edited by Gerhard Preyer and Georg Peter. New York, NY: Oxford University Press.

Borg, Emma. 2009. Semantic Minimalism. In *The Routledge Pragmatics Encyclopedia*, edited by Louise Cummings. London: Routledge.

Borg, Emma. 2010. Minimalism and the Content of the Lexicon. In *Meaning and Context*. Edited by Luca Baptista and Erich Rach. Bern: Peter Lang.

Borg, Emma. 2012. *Pursuing Meaning*. Oxford: Oxford University Press.

Brown, Susan W. 2008. Polysemy in the Mental Lexicon. *Colorado Research in Linguistics* 21: 1-12.

Cappelen, Herman and Lepore, Ernest. 2005. *Insensitive Semantics: A Defense of*

---

copo Tagliabue for valuable input on an earlier draft of the paper. I also thank two anonymous reviewers for their comments on the submitted manuscript. The usual disclaimer applies. Part of the research that led to this paper has received funding from the European Union's Seventh Framework Program (FP7/2007-2013, MSCA-COFUND) under grant agreement No. 245743, Post-Doctoral Program Braudel-IFER-FMSH, in collaboration with the Institut Jean Nicod and the Labex IEC, ENS Paris.

Semantic Minimalism and Speech Act Pluralism. London: Wiley.
Chomsky, Noam. 2002. *On Nature and Language*. Cambridge: Cambridge University Press.
Coltheart, Max. 1999. Modularity and Cognition. *Trends in Cognitive Sciences* 3: 115-120.
Cover, Thomas M. and Thomas, Joy A. 2006. *Elements of Information Theory*. 2nd edn. London: Wiley.
Davis, Wayne A. 2013. On Nonindexical Contextualism. *Philosophical Studies* 163: 561-574.
Devitt, Michael. 2013. Is There a Place for Truth-Conditional Pragmatics? *Teorema* XXXII/2: 85-102.
Evans, Vyvyan. 2009. *How Words Mean: Lexical Concepts, Cognitive Models, and Meaning Construction*. Oxford: Oxford University Press.
Fodor, Jerry A. 1983. *The Modularity of Mind*. Cambridge, MA: MIT Press.
Folk, Jocelyn R. and Morris, Robin K. 2003. Effects of Syntactic Category Assignment on Lexical Ambiguity Resolution in Reading: An Eye Movement Analysis. *Memory & Cognition* 31: 87-99.
Frazier, Lyn and Rayner, Keith. 1990. Taking on Semantic Commitments: Processing Multiple Meanings vs. Multiple Senses. *Journal of Memory and Language* 29: 181-200.
Giora, Rachel. 2003. *On Our Mind: Salience, Context, and Figurative Language*. New York, NY: Oxford University Press.
Giora, Rachel. 2012. The Psychology of Utterance Processing: Context vs Salience. In *The Cambridge Handbook of Pragmatics*. Edited by Kasia M. Jaszczolt and Keith Allan. Cambridge: Cambridge University Press.
Harley, Trevor A. 2014. *The Psychology of Language: From Data to Theory*, 4th edn. New York, NY: Psychology Press.
Heim, Irene. 1983. On the Projection Problem for Presuppositions. In *Second Annual West Coast Conference on Formal Linguistics*. Edited by Dan Flickinger and Michael Westcoat. Stanford, CA: Stanford University Press.
Heim, Irene. 1988. *The Semantics of Definite and Indefinite Noun Phrases*. New York, NY: Garland.
Jarema, Gonia and Libben, Gary. Eds. 2007. *The Mental Lexicon: Core Perspectives*. Amsterdam: Elsevier.
Jaszczolt, Kasia M. 2007. On Being Post-Gricean. In *Interpreting Utterances: Pragmatics and Its Interfaces*. Edited by Randi A. Nilsen, Nana A. A. Amfo and Kaja Borthen. Oslo: Novus.
Kamp, Hans and Reyle, Uwe. 1993. *From Discourse to Logic*. Dordrecht: Kluwer.
Kennedy, Christopher. 2007. Vagueness and Grammar: The Semantics of Relative and Absolute Gradable Adjectives. *Linguistics and Philosophy* 30: 1-45.
Kennedy, Christopher and McNally, Louise. 2010. Color, Context and Compositionality. *Synthese* 174: 79-98.
Klein, Devorah E. and Murphy, Gregory L. 2001. The Representation of Polysemous Words. *Journal of Memory and Language* 45: 259-282.
Klepousniotou, Ekaterini. 2002. The Processing of Lexical Ambiguity:

Homonymy and Polysemy in the Mental Lexicon. *Brain and Language* 81: 205-223.
Korta, Kepa and Perry, John. 2006. Varieties of Minimalist Semantics. *Philosophy and Phenomenological Research* 73: 451-459.
Levinson, Stephen J. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature.* Cambridge, MA: MIT Press.
MacFarlane, John. 2007. Semantic Minimalism and Nonindexical Contextualism. In *Context-Sensitivity and Semantic Minimalism: New Essays on Semantics and Pragmatics.* Edited by Gerhard Preyer and Georg Peter. New York, NY: Oxford University Press.
MacFarlane, John. 2009. Nonindexical Contextualism. *Synthese* 166: 231-250.
Manning, Christopher D. and Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.
Mason, Robert A. and Just, Marcel A. 2007. Lexical Ambiguity in Sentence Comprehension. *Brain Research* 1146: 115-127.
Matsuki, Kazunaga, Chow, Tracy, Hare, Mary, Elman, Jeffrey L., Scheepers, Christoph and McRae, Ken. 2011. Event-based Plausibility Immediately Influences On-Line Language Comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 37: 913-934.
Morris, Robin K. 2006. Lexical Processing and Sentence Context Effects. In *Handbook of Psycholinguistics*. Edited by Matthew J. Traxler and Morton. A. Gernsbacher, 2nd edn. San Diego, CA: Academic Press.
Murphy, M. Lynne. 2010. *Lexical Meaning.* Cambridge: Cambridge University Press.
Pelczar, Michael. 2000. Wittgensteinian Semantics. *Noûs* 34: 483-516.
Piantadosi, Steven T., Tily, Harry and Gibson, Edward. 2012. The Communicative Function of Ambiguity in Language. *Cognition* 122: 280-291.
Pinker, Steven and Bloom, Paul. 1990. Natural Language and Natural Selection. *Behavioral and Brain Sciences* 13: 707-784.
Recanati, François. 2004. *Literal Meaning.* Cambridge: Cambridge University Press.
Searle, John. 1980. The Background of Meaning. In *Speech Act Theory and Pragmatics*. Edited by John Searle, Ferenc Kiefer and Manfred Bierwisch. Dordrecht: Reidel.
Seidenberg, Mark S., Tanenhaus, Michael K., Leiman, James M. and Bienkowski, Marie. 1982. Automatic Access of the Meanings of Ambiguous Words in Context: Some Limitations of Knowledge-Based Processing. *Cognitive Psychology* 14: 489-573.
Small, Steven, Cottrell, Garrison, and Tanenhaus, Michael K. Eds. 1988. *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*. San Mateo, CA: Morgan Kaufman.
Spivey, Michael, McRae, Ken, and Joanisse, Marc. Eds. 2012. *The Cambridge Handbook of Psycholinguistics*. Cambridge: Cambridge University Press.
Stojanovic, Isidora. 2008. The Scope and the Subtleties of the Contextualism/Literalism/Relativism Debate. *Language and Linguistics Compass* 2: 1171-1188.
Swinney, David A. 1979. Lexical Access During Sentence Comprehension: (Re) Consideration of Context Effects. *Journal of Verbal Learning and Verbal Behavior*

18: 645-659.
Tanenhaus, Michael K., Leiman, James and Seidenberg, Mark. 1979. Evidence for Multiple Stages in the Processing of Ambiguous Words in Syntactic Contexts. *Journal of Verbal Learning and Verbal Behavior* 18: 427-440.
Traxler, Matthew J. 2011. *Introduction to Psycholinguistics: Understanding Language Science*. London: Wiley-Blackwell.
Van der Sandt, Rob. 1992. Presupposition Projection as Anaphora Resolution. *Journal of Semantics* 9: 333-377.

Defending Backwards Causation against the Objection from the Ignorance Condition

Abla Hasan
University of Nebraska-Lincoln

BIBLID [0873-626X (2014) 39; pp. 173-197]

Abstract
Since Michel Dummett published "Can an effect precede its cause?" (1954), in which he argued for the logical consistency of backwards causation, the controversial concept has turned to a subject of all kinds of interpretations and misinterpretations. Some like Ben-yami, Peijnenburg and Gorovitz have wrongly ascribed to Dummett the view that the argument for the consistency of believing in backwards causation applies only in cases where the agent doesn't know about the occurrence of the past effect. In this paper I defend Dummett's argument by clearing up the confusion caused by ascribing the ignorance condition to Dummett.

Keywords
Cause, effect, backwards causation, ignorance condition, logical consistency

First, let me explain what I mean by the ignorance condition. The ignorance condition is the thesis that the agent needs to be ignorant of the occurrence of the past event to have a good reason to consistently believe that bringing about that event by doing something in present makes sense.

Ben-Yami in his paper "The impossibility of backwards causation" argues that Dummett's condition can't be fulfilled and consequentially makes the argument for backwards causation impossible. Ben-Yami says,

Disputatio, Vol. VI, No. 39, November 2014
Received: 17/12/2013 Revised: 16/06/2014 Accepted: 02/09/2014

> Dummett acknowledged the difficulty which the bilking argument creates for the possibility of backwards agent-causation. He presents the argument and tries to reply to it on pp. 352-8 of "Bringing about the Past". His conclusion is that for backwards agent-causation "to make sense" the agents should be incapable of knowing whether the earlier event, which they are now trying to cause, has occurred (Ben-Yami 2007: 446)

Therefore, in his argument against the coherence of backwards causation Ben-Yami uses cases where the agent's knowledge of the occurrence of the past event seems to be separable from the agent's intention to bring about that type of event. He says, "specifically, we need to assume that any alleged cause can be prevented if agents know that b [the young men have been brave] has occurred" (Ben-Yami 2007: 443).

He puts his conclusion as the following

> I have thus shown that one of the conditions Dummett finds necessary for an agent's action to produce an effect, namely, that the agent cannot know, at the time of the action, whether the effect has occurred, makes it impossible for the case to be one of backwards causation. (Ben-Yami 2007: 455)

The same view, that ascribes to Dummett the belief in the ignorance condition as a condition for the consistency of backwards causation, appears also in Jeanne Peijenburg's paper "Shaping your own life". She says, "Dummett's argument implies that something did happen in the past, and on condition that we do not know what it is, we are able to exert some influence" (Peijnenburg 2006: 245). Similarly, we find Gorovitz arguing for the agent's ignorance as a condition for Dummett's argument. He states,

> In his earlier article [Dummett] on reverse causality, he suggests that the causal connections that work in reverse are effective only in cases where the agent is ignorant of whether or not the desired event has actually occurred. (Gorovitz 1964: 369)

However, this condition, i.e., the ignorance condition, is not required by Dummett in the first place and ascribing this condition to Dummett is a misinterpretation of his argument. Therefore in this paper I clarify the misinterpretation by explaining Dummett's position on the ignorance condition. I argue that, for Dummett, the agent's ignorance of the occurrence of past events is not a condition

for the consistency of believing that the causes for those past events can be located in the present. Therefore, reading Dummett as requiring the ignorance condition as a condition for his argument for the logical consistency of backwards causation is a misinterpretation.

But first, and in order to be able to fully appreciate Dummett's view on the ignorance condition, we need to understand what both the concept of the past and the concept of the future mean for Dummett; more importantly speaking, what do our knowledge of the past and our knowledge of the future refer to? According to Dummett, we have two kinds of knowledge: knowledge of the future and knowledge of the past. Our knowledge of the future can be analyzed into: "prediction based on causal laws and knowledge in intention" (Dummett 1980: 344). Our knowledge of the future as based on prediction can be something like my knowledge that it might rain tomorrow, as I observe the cloudy sky this evening. This kind of knowledge is simple and based on causal laws. In addition, there is my knowledge of the future as based on intention; for example I know that the door will be opened after five minutes because I intend to open the door after five minutes. On the other hand, Dummett believes our knowledge of the past to be based on our memories and on deductions from the present. Dummett says, "with our knowledge about what has happened in the past, it is quite different: we have our memories, and we also have deductions from what is the case now, based upon our belief in certain causal regularities" (Dummett 1980: 331). I might for example know that my precious vase was broken yesterday because I remember seeing it falling on the floor and smashing into pieces; this kind of knowledge is based on my own observation, which can be preserved by the aid of memory. But in addition, I might know things about the past without personally observing them taking place in the past; this is what Dummett refers to as the knowledge of the past as based on deductions from the present: here comes the role of the causal laws again, because I can know for example that it has rained yesterday from observing the wet yard as I wake up in the morning and without really the need to observe it raining.

What Dummett rightly observes is that knowledge in intention has no parallel in the case of our knowledge of the past, because while we take knowledge in intention as a sufficient way to know about fu-

ture we don't similarly appeal to our knowledge in intention when it comes to our knowledge of the past; this takes place mainly because we are beings that have memory but don't have foreknowledge. I don't have to appeal to my knowledge in intention as applied to past events the way I appeal to it when it comes to future events, because I can always know about the occurrence of past events via what my memory tells me in cases of my own observations, and what others' memories can tell me in cases of depending on others' testimony.

But what if we had foreknowledge? Dummett asks; then, the picture would be the same for both cases of future as well as for the past. Then we would not trust our knowledge in intention as directed to the future the way we trust it now, and we would prefer knowledge based on our foreknowledge faculty the way we prefer now knowledge based on our memory. Then it would be difficult to form intentions directed to the future in the way we now find it difficult to form intentions directed to the past. Dummett says, "For us to be able to form future intentions at all, we should have to have a cognitive attitude to the future not wholly analogous to our present attitude to the past" (Dummett 1996: 362).

This difference, i.e., between having memory and not having foreknowledge, is of great importance when it comes to the way we form our causal beliefs towards the future or towards the past. We have a special faculty that enables us to know what has happened and to say for sure that what has happened has happened: this faculty is memory. But when it comes to the future we don't have, as human beings, a similar faculty that we can trust to inform us about what will happen in the way we can trust what our memories tell us about what has happened. Now, the consequence of not having foreknowledge but having memory is that we don't feel that knowledge in intention can be applicable in the case of knowing the past in the way it can be applicable in the case of knowing the future, because we have ways to know about the past other than via knowledge in intention. Therefore, when we are informed about the nonoccurrence of past events by the ordinary ways of being informed, we usually take their nonoccurrence for granted even if those types of events are in our power to bring about by doing something in the present.

In other words, if the agent is informed by the ordinary ways of being informed about the nonoccurrence of a past event and he had

the present intention to bring about that event which is in his power to be brought about, he will take the nonoccurrence of that past event as granted but he will take a different attitude if he is informed about the nonoccurrence of a future event that he intends to bring about by doing something, because he will trust his knowledge in intention in the case of the future but will hesitate to do the same in the case of the past event.

Let's take a simple example. Suppose that I have recently moved with my family to a new house. The landlord provides us with some keys but then he warns us that some of those keys might not really work properly. As he gives us all three keys, he tells us that he can't really identify which ones might not be working because he doesn't have the time to try all of them and because we can simply know by way of trying which ones do not work. However, I get my key and I give the other two to my husband and my son. Suppose that like our landlord we didn't have the time to try the three different keys we have got. But as I return from work and I try my key and it works perfectly, I know that mine is working. For a couple of days I keep using my key to open the door with no problems at all. The next week my son calls me to say that his key which he is using for the first time is not working and he tells me furthermore that we need to call the landlord because the door can never be opened unless he provides us with extra keys. What would be my response in this case? Simply, I will assure my son that I am on the way and that there is no need to call the landlord, because I have my keys. Now, the question is how do I know that the door will be opened? The answer is because of my knowledge in intention as applied to the future.

I can know and even assure my son that the door will be opened after I arrive, because I have the intention to open the door by using my key when I arrive. In other words, my belief in the opening of the front door lies in my intention to open the front door. Let's suppose that my son called me on the phone to tell me that he came home yesterday and couldn't open the front door by using his key. In this case I don't have any way to know about what has happened other than my son's testimony; because his testimony provides me with knowledge based on his memory. Appealing to knowledge in intention is not even an option when we have knowledge based on memory, because we have memory but we don't have foreknowledge. By depending on

our memory — excluding cases of hallucinations and being unable to recall what has taken place — we believe we can know what has happened for sure, and therefore appealing to knowledge in intention seems superfluous and even awkward. In the previous example, when my son tells me that he couldn't open the door yesterday, I take the truth of what he is saying for granted. His testimony would be in this case the most reliable way to know what has happened when he tried to open the door. But, when he calls me to tell me that the door can't be opened today, I can appeal to my intention to use my key to open the door to believe that, contrary to what my son is reporting, the door will be opened today, because I intend to open the door after arriving.

In the previous example, I know that the door will be opened because I believe that I have the working keys and this can be sufficient to open the door. But let's suppose that I arrived at the house and this time, unlike all the other times I tried my keys, the door didn't open. Suppose that the reason for the door not to be opened was different from what both my son and I have suspected. Suppose that both our keys are working, but, because of my son's wrong attempts to open the door, the lock of the door has been damaged from the inside. In this case, although I had a reason to believe that the door will be opened because of my intention to open the door, my knowledge in intention, one way I have to know about the future, was proven not to be as reliable as our ordinary ways of knowing about the past.

In short, we apply knowledge in intention to know the future because we have memory but we don't have foreknowledge. The reliability of the other ways we have of knowing about the past makes us eliminate knowledge in intention as a way to know the past. However this doesn't lead Dummett to eliminate knowledge in intention as a considerable way to think of the past under certain conditions that he describes. Before we go through those conditions, what seems of a great importance in order to be able to understand Dummett's position on the ignorance condition is to distinguish two ways of being informed about the occurrence of events in the future or in the past. First, there is knowledge in intention; and second, there is knowledge as acquired by ordinary ways of being informed. Now, in the case of the future, ordinary ways of being informed in addition to knowledge in intention are knowledge as based on predications

from the present, while in the case of the past, ordinary ways of being informed are deductions from the present and what memory can tell us.

Now, the important step that Dummett takes is viewing knowledge in intention as being applicable in principle to the past in the way it can be applicable to the future.

Consider an agent who:

1. Observes a systematic repetition and regularity between doing a type of act A in the present and observing the occurrence of a type of event E in the past.

2. Can find no causal explanation for that type of past effect E by going back to previously occurring events prior to E.

3. Can find no ordinary causal relation between the past type of event E and the present type of act A that proves the type of event E to be the cause of the of act of type A.

4. Can find no incidents in which he tries to do A but he fails, i.e., type A of action is always in his power to be performed.

Such an agent will have a good reason for believing that his intention to bring about the type E of event, by way of performing type A of act, makes sense. But more importantly, such an agent will have a reason to suspect any informant that tells him about the nonoccurrence of the past type of event E when he has the intention to bring it about by performing A. But this result has a specific importance when it comes to the ignorance condition, because the agent in this case can't be said to be ignorant of the occurrence of the past type of event E; in fact, he can't be ignorant of the occurrence of the past type of event E, because the occurrence of the past type of event E will fall under his knowledge in intention, which he will trust more that the ordinary ways of being informed about the past if the previously mentioned conditions are fulfilled.

Suppose that for the last four years I observed that every time I make a phone call from the department, I hear about a heavy snowing in my country that starts exactly two hours before the time I made my call. Suppose that being aware of the awkwardness of my

belief I tried everything to make sure that I have a reason to believe that it is really my calling from the department what is causing the previous snowing, even in times when it is not really supposed to snow there. After trying almost all possible ways to check my belief, I arrive, after four years, at a point where I believe that making a phone call from the department causes a previous snowing. I try all Dummett's conditions and I find a perfect match:

> First, calling is always proven to be something that I can do and there are no cases where I try to make the phone call but I couldn't.
>
> Second, I couldn't find any other ordinary causal explanation for the snowing in times and places where snowing is not even expected.
>
> Third, there is a systematic repetition and regularity between my calls and the snowing.
>
> Fourth, snowing in my country can't be explained as the cause of my calls.

In this case, I would have a reason to believe that I am causing it to snow 2 hours before I make phone calls by way of making those phone calls. Furthermore, after I reach my belief in the consistency of trusting my backward causal ability to affect the past, and after checking the fulfillment of the four conditions, I can't even be considered to be ignorant of the past occurrence or nonoccurrence of the past snowing. Because even in cases where ordinary ways of being informed about the past events are not available, I can still appeal to my knowledge in intention as directed to the past. I can simply claim that I have the ability to know about the snowing before I hear about that from any one, because my intention to bring about that snowing by way of making a phone call makes me no more ignorant of the snowing. Of course, what I am claiming sounds strange, but remember that this is something that I tried for four years to check and it was proven to be working all the time. I might even start using my discovered new ability to call my friends back home and tell them about the snow even before they say anything.

In short, in cases where the agent establishes a backward causal belief about the past, the agent can't really be said to be ignorant about the occurrence of the past event, since the concurrence of that past event lies in the agent's knowledge in intention as applied to the past.

But let's suppose that an agent, whom I will refer to as 'Dummett's agent', was informed about the nonoccurrence of the past event which he intends to bring about. Would that motivate him to give up his attempt to bring about that event? No, because in this case Dummett's agent will take the position of any ordinary agent who intends to bring about some effect in the future and is informed about the nonoccurrence of that event. He will simply suspect the information he has got about the nonoccurrence of the event in question, instead of suspecting his knowledge in intention.

Dummett's agent who is informed about the nonoccurrence of the past event that he intends to bring about will not lose his motivation to bring about that event; rather, he will interpret the information as being false and not trustworthy, simply because the occurrence of that type of event lies in his knowledge in intention.

In "Bringing about the past" Dummett says about an agent who is informed of the nonoccurrence of a past event E which he thinks to be the effect of a present action A,

> Now he need not really deny that learning, in the ordinary way, that E has not occurred makes it at all more probable that, if he tries to perform A he will fail. He may concede that it makes it to some extent more probable, while at the same time maintaining that, even when he has grounds for thinking that E has not occurred, his intention to perform A still makes it more probable than it would otherwise be that E has in fact occurred. The attitude of such a man seems paradoxical and unnatural, but I can't see any rational considerations which would force him out of this position. (Dummett 1980: 349)

But what if the agent was not merely told about the nonoccurrence of the past event in question? What if he has seen or experienced the nonoccurrence of the past event himself; would that change the case? In other words, what if we replaced others' testimony by memory as another way of knowing the past? Would this kind of replacement change the argument in favor of allowing at least a weaker version of the ignorance condition? A version that requires the agent to have no "memory-based" access to the occurrence or the nonoccurrence of

the past event to rationally believe in bringing about that past event by way of doing something in the present. To anticipate what will be presented, the answer to this possibility is no. I argue that the agent who depends on memory as one way to objectively know about the past, similar to the agent who depends on others' testimony, doesn't need to be ignorant — either partially or fully — about the past to believe in the consistency of backwards causation. To explain more, let's consider the case that Dummett suggested as his typical faithful believer in backwards causation, or what is known as "the dancing chief" (Dummett 1980: 343).

In "Bringing about the past", Dummett asks us to imagine a tribe that has a specific custom; every second year the young men of the tribe are sent on a lion hunt to prove their manhood; during this ritual they travel for two days, hunt lions for two days, and spend two days on the return journey. Observers accompany them in their trip to report to the chief upon their return whether the young men were brave or not. The tribe's whole causal beliefs are different from ours; they hold that some ceremonies performed by the chief have the ability to influence weather, etc; but what is important to be remembered is that these ceremonies are not to be taken as related to gods of any kind at all. Now, while the young men are away from the village, the chief performs ceremonious dances intended to cause the young men to act bravely.

Let's suppose that the chief continues to perform these dances for the whole six days that the party is away. His act can be considered as a case of an act performed for the purpose of bringing about the past.

Now, let's apply the distinction between memory and others' testimony, as two distinguished ways to know about the past, to examine whether the distinction will require the agent at least to have no "memory-based" knowledge of the occurrence of the past event to have a rationale for believing that he can bring about that type of past event by causing its causes in the present. But first, let's reinvestigate what would knowledge of the past as based on others' testimony be like in this case. As I mentioned earlier, an agent who has a belief based on a long experience, and fulfills all the conditions specified by Dummett for holding a backward causal belief, doesn't depend any more on other ordinary ways of being informed about the past that go beyond his knowledge in intention. As Dummett explains, the

chief already believes on the basis of his past experience that his act of dancing is able to bring about the previous bravery of the young men; he is not testing a new causal hypothesis by trying to dance and then watching what will happen, or what the reports would tell him about what had happened.

Now, whether the chief is justified in holding such a belief or not, or even what grounds he has for holding it is another issue that is of no importance here. What matters is that what really causes the chief to dance during this particular hunting trip made by the young men is his belief that his dancing will cause the young men's previous brave behavior.

What should be remembered here is the fact that he is not trying to test what would his dance accomplish or how can his dance causally affect the past behavior of the young men; rather, he is already convinced that there is a causal connection between his dancing and their past brave behavior. This means that Dummett's chief, already convinced of his backward causal powers, will simply interpret any reports of the young men not being brave as false, since his rationale not to believe those reports after he danced successfully as usual and as he did — let's say for the last 20 years — will be stronger than the rationale to believe such kind of reports. The case of ignorance as based on others' testimony is then ruled out. The bravery of the young men is a knowledge already acquired by the chief as he dances to bring that previous bravery about.

However, what about knowledge as based on memory? Does it form a case different from the previous case? Practically speaking, what if the chief decides to accompany the young men in their hunting trip instead of waiting patiently for the reporters to come back and tell him about what had happened? Would the case be any different? Would the chief lose his rationale in believing in his ability to cause past brave behavior among the young men, simply because of observing the young men acting differently? On the opposite, the chief who accompanies the young men won't be that different from the chief who waits for the reports to be sent to him. Because while the second will interpret what is said to him as false, the first will interpret what he observes by himself as false. He might for example interpret what he observes as mere hallucinations, day dreams, tricks played by the young men, etc, simply because in both cases the

chief's deeply rooted belief, as based on his long experience, goes beyond what can be falsified by whatever can be said or even observed.

The chief's belief in such a strong causal connection that doesn't seem to need much further investigation is not that odd, although it might look like that. It is a common feature of our epistemic system, because we, as human beings, don't repeatedly reinvestigate our already established causal beliefs; rather, we just take them for granted. In fact, any advancement in our human knowledge would be impossible if reinvestigating all causal beliefs that we assert is required each time we assert them. Our everyday behavior tells a different story; we investigate our causal beliefs only as we are establishing them, but when they are already established, we only apply them. For example, the teacher who wants to explain the law of gravity to his students will simply drop the pen from his hands, not to see where the pen will go, or to examine if the pen will go to the ground or not, but to show his students that the pen will certainly land on the ground in seconds. The teacher in this case is not interested in examining his causal belief as much as he is interested in demonstrating it; he might not even feel the need to look at the pen while it is falling to the ground; he might simply drop the pen and turn his face to his students.[1] In short, already existing causal beliefs, when transferred into new cases, don't depend on observation, because observation takes place only when the causal belief is still to be established. However, it is very important for my argument to keep in mind that what I mean by casual connections, are those causal connections taking place between types of events and not simply causal connections taking place between events.

However, this has been addressed early in the four conditions required of an agent to rationally believe in the consistency of backwards causation. Therefore, what matters in the case is the belief that dropping things will cause them to fall down, and not the concrete individual case of dropping a pen. Therefore, the first instance of a case to opposite, where the teacher fails to cause the pen to fall to the ground by dropping it, won't immediately motivate him to move from the mental statues of transferring already existent causal

[1] As I made clear earlier I am limiting my argument to cases in which the cause is the necessary as well as the sufficient reason for bringing about the effect.

beliefs to establishing new causal beliefs; simply because what matters are not events but types of events. The clash between the already established causal belief that the teacher holds and the temporary impossibility — that he observes for the first time — is not to be understood as a clash between two causal beliefs per se, and better to be understood as a clash between a causal belief, already established between types of events, and a causal connection to the contrary between two events; in other words, the teacher's observation to the causal connection between throwing the pen and the pen's floating in the air will not be evaluated by him the same way he evaluates the causal connection between throwing the pen and the pen's falling down to the ground. Simply, because the first case mentally represents a mere connection between events, while the second represents an already established causal belief based on a connection between types of events. One can easily predict that it will be a long way before this isolated causal connection, between the two events of throwing the pen and its floating in the air, can be mentally affirmed and established, and only then it can become transformable to new cases.

In fact, establishing causal beliefs is an ongoing process that only starts by events and doesn't fulfil until it forms a belief in a casual connection between types of events. Similarly, the chief who believes in his ability of bringing about a past bravery of the young men had to go through a long process of repeatedly experiencing an actual ability of bringing about past bravery by way of dancing over the years. As a result, the chief's first response to the denial of his long term causal belief — as taking place between types of events — will be to deny what the observation from one event says.

Of course, my argument doesn't mean the total negation of any role of observation after the causal belief is established. As a matter of fact, repeated cases of counter examples can eventually motivate one to suspect the reliability of his/her already established causal beliefs. In the long run, this suspension and reevaluation of one's causal beliefs can end by forming contrary new causal beliefs. However, what is important to keep in mind, is the fact that forming the new causal beliefs would require going through a totally new process of establishing those new opposite causal beliefs.

In all cases, as we have seen, observation doesn't play much of

a role in applying previously established causal beliefs, because it is taken for granted. It starts to play a role only if it repeatedly starts giving results opposite to what is expected, and even in this case, the process of suspecting the already established causal belief and forming a new causal belief is a long process that a single observation can't suffice to bring about.

How does that affect the ignorance condition? Well, as I argued before, observation is an essential step for establishing new causal connections; more precisely speaking, for the mental affirmation of the existence of causal connections. But after the causal connection is established, observation is no more needed, at least in the way we understand causal connections and deal with them. As with the case of the teacher, the chief who already believes in a causal connection between his present dancing and the past brave behavior of the young men doesn't even need to bother himself to make any effort to know whether the young men have been brave or not, because what matters for him is not the occurrence of the effect itself, but the soundness of the whole causal belief he had. This is because the act of his dancing is not an act made for the purpose of *discovering or establishing a* new causal connection, on the contrary, it is an act of *applying* an already affirmed causal connection to other cases.

One might object that the ignorance condition applies before and not after the chief dances. It applies when the chief knows that the young men have been brave before he dances or when he knows that the young men have not been brave; consequentially, what we really need is a discussion of the chief's motivation and rationale to dance after he knows about previous bravery, and before he performs any dances. Now, the two cases can be easily evaluated if we avoid a confusion that might take place here between two causal beliefs,

(a) if I dance I can cause previous bravery of the young men.

(b) if I dance I can cause the bravery of the young men.

The chief's basic causal belief is (a) and not simply (b). For example, in the case where the chief knows about the previous bravery of the young men, what he is informed of is the previous bravery which he believes to take place because of later dancing, and not simply because of dancing. Therefore, even after knowing that the young men have

been brave, he will still be motivated to dance, because what he believes in is the effectiveness of his later dancing, and not merely his dancing. The chief who discovers that the young men have been brave, even before he dances, would still be motivated to dance, because — according to his belief — it is his later dancing what really causes the previous bravery of the young men. Here, it is always important to remember that according to the chief's causal belief, the bravery of the young men is a previous effect of a lately occurring cause.

Let 'B' stand for the young men's being brave and 'D' stand for the chief's dancing. Let's suppose that the chief was able to know about the brave behavior of the young men in one way or another; let's say that one reporter stood on a top of a mountain and signaled to him informing him of the bravery of the young men even before he started dancing. How can one explain what will take place in this case? When the chief is informed of the bravery of the young men what he is informed of is the previous bravery, while his belief which forms the real cause for his act of dancing is the causal belief that later D causes previous B and not merely previous B; so when the chief is asked for example by his grandson, "why do you dance after you have already been informed of the bravery of the young men?" he will simply say, "because my dance is the cause of the bravery of the young men". In other words, he will refer in his answer to his belief in the causal law; he will not refer only to B, because basically it is his belief in the causal law what motivates him to dance. But, if his grandson was still curious and he asked him another question, "but why do you still have to dance while you know that the young men have already proven their bravery, aren't you wasting your time by doing that?" Probably the chief would answer him by saying, "Well, because it is this dance that I am going to perform that has caused them to be brave, if they have really been brave. If the young men have been brave they must have been brave only because I will dance now."

To draw an analogy with the future; suppose that the chief, instead of believing that his later dance causes the previous bravery of the young men, believes that drinking a special drug can keep him from getting sick. Suppose the chief — in an attempt to stay healthy — makes sure to drink this drug every day before breakfast. Now, let's imagine that one day his curious grandson asks him the question: "why do you take that drug every day?" Then the chief would

probably say something like: "because I want to be healthy". What if the curious grandson was not satisfied and insisted on getting a better answer by saying: "But you are healthy!"? At this point of the dialogue, the chief would probably say something similar to what he says in the case of defending his dancing, because he might say something like: "but taking the drug is what makes me healthy".

Actually, changing causal beliefs after they are already established is not as easy as some might think, because a new process of forming the opposite causal beliefs would be needed before one can be able to assert them. I argued that the chief who is informed of the occurrence or the non-occurrence of the previous effect will still have a rationale to believe that he can explain the previous effect by later causes; furthermore the chief who after dancing is informed of the non-occurrence of the previous effect has a rationale to suspect what he is told or even what he observes, because he will have two contradictory beliefs, first: his belief in the law that later D causes previous B, and second the belief that previous B did not occur even though later D occurred. Therefore, he must abandon one of the two beliefs.[2]

The chief's belief in the causal law if already established from the past — as is the case with any causal belief — will no longer need to appeal to any observation to reestablish it; if the chief was sure of the truth of the law, then he would hesitate not to dance even after hearing reports about B because, if the previous bravery was the outcome of his later dancing, then not dancing after he is informed of their bravery will end up in one way or another in the young men's not having been brave. He might think that, if their bravery is caused by his later dance, and he was sure of that causal relation, and he didn't dance after he hears about their previous bravery, the consequence will be not believing that the young men were brave, because if the causal belief is what provides him with the rationale for believing that the young men have been brave, not dancing will leave him without any reason for believing in the occurrence of B. It might be the case that he doesn't dance and the reporters tell him, after coming back from the trip, that the young men have been brave, but he might sim-

[2] Here I am considering the case where later D is a sufficient as well as a necessary cause of B, to eliminate the complexity from having other equally effective causes.

ply not believe the reporters; he might think that the reason he has not to believe the reporters, namely, not dancing, is stronger than the reason he has to believe them, namely, their testimony.

In other words, in case the chief didn't dance after he has been informed about the previous behavior of the young men, he will start to suspect B itself instead of suspecting the causal law, because as I have said before, the chief is 100% sure of his belief in the causal law 'later D causes previous B'. The observation that B is not enough for him to suspect the causal law; consequentially, he will be left only with one option, namely, to suspect the occurrence of B. Similarly, our confident chief will not be motivated to suspect the occurrence of the later dancing as he witnesses a cowardly behavior of the young men, the same way he doesn't suspect the previous bravery; for example, he will not say to himself as he observes the cowardly behavior of the young men, "probably I will not be able to dance this time, probably I will slip on a banana peel"; this is because his observation is merely based on one event, while his causal belief connects types of events and not events. In other words, the chief will have a rationale to suspect his ability of dancing if his belief was 'This later dance D will case previous bravery B'. But his belief is more like: 'Later D causes previous B' or more accurately speaking, his belief is more like: 'Later type of D causes previous type of B.'

In his objection to Dummett's argument in favor of the consistency of backwards causation, Gorovitz in his paper "Leaving the past alone" wrongly — as I discussed before — ascribed to Dummett the agent's ignorance of the occurrence of the past event as a condition for the consistency of his believing in backwards causation. But furthermore, he discussed a case he claimed that Dummett omitted. It is the case where the chief and instead of waiting for the reporters to come back and inform him about the behavior of the young men insists on witnessing the hunt himself. In this case Gorovitz concludes no dance will be necessary. He asserts this by saying, "I conclude that if the chief witnesses the warriors being brave, no dance is necessary" (Gorovitz 1964: 368).

He describes this case by saying,

> In Dummett's example, various experiments are described that are designed to show the chief to be in error in his beliefs. These experiments all fail. But there are a few more which Dummett did not consider. For

> example, instead of letting the chief remain at home while the warriors hunt, only to cause after their return their having been brave on the hunt, let us insist that the chief himself witnesses the hunt. Now there can no longer be any question of lying reporters who try to deceive the chief, only to be discovered when he dances. The chief will himself observe the cowardice or bravery of the men. Then if they were brave, he will of course have no need to dance. (Gorovitz 1964: 368)

But taking into consideration my distinction between transferring already established causal beliefs and establishing new causal beliefs we can find a way to answer the case designed by Gorovitz. Because the chief who already believes that his later dancing always causes previous bravery of the young men will continue trusting his always-trusted belief even if he witnessed the hunt himself. In fact, the difference in the way he gets the opposite information to what he believes in will not be immediately trusted by him, his being informed about the cowardly behavior of the young men before he starts dancing or even his witnessing himself the cowardly behavior before he starts dancing. This will not shake his belief in the causal law, because the chief who already believes that his later dance causes previous brave behavior is not even depending anymore on what he is informed about, or what he witnesses himself to confirm his beliefs. In both cases he will find a way to interpret what he is informed about or what he sees in way that fits his causal beliefs. The stronger and better established his beliefs are, the less the chief will take what he sees or hears about to be reliable enough to change his beliefs. The chief who believes that his later dance causes previous bravery will simply interpret all reports about the non-brave behavior of the young men as being false in case he was informed about that after he dances, but in case he was told about the non-brave behavior of the young men before he dances, this will enforce his belief that the bravery of the young men is conditioned by his dancing. Similarly, Gorovitz's chief, who, unlike Dummett's chief who stays home and waits for the reporters to come back, insists on witnessing the hunt himself, will interpret the behavior he witnesses either cowardly or brave, in a way that fits his causal beliefs. This means in case the chief didn't dance yet and he witnesses a cowardly behavior of the young men, this will insure his belief that he should have danced to make them behave bravely, and his belief will be asserted; but in case the

chief witnesses a brave behavior of the young men before he dances, he will simply interpret that as false. He will accuse the young men as acting in a way to deceive him or being involved in a conspiracy against him, or he might interpret what he witnesses as false and only as hallucinations because of his getting old. In short, Gorovitz's dancing chief is not that different from Dummett's chief, and the case Gorovitz designs doesn't necessarily end as he claims by causing the chief not to dance.

To sum up what has been said until now, the chief's rationale for dancing is his belief in the causal law: later D causes previous B; this causal belief, like any other causal belief that one might have, is formed according to previous observation or previous set of observations of a regularity that takes place between previous B and later D, or it might be based on testimony, because as said before, it might be the case that the act of dancing to bring about the previous bravery of the young men is a tradition that the chief inherited from his father, who inherited it from his father, and so on. What matters is that after the causal belief is already formed, the human mind usually moves to another step, which is applying this same already existing belief as constructed in the past to new cases in the present and even in the future, without much investigation.

In the case of the dancing chief, if he has any causal belief at all, he might have one of the two: either later D causes previous B, or later D causes ~ previous B, according to what he had concluded from his previous observations. Now, if the chief believes that later D causes previous B, then his knowledge of the occurrence of later B will not affect his act of dancing, because he thinks that the occurrence of previous B is caused by the later act of dancing that he will perform. Therefore, the occurrence of previous B will be interpreted by him as being not true if it was not followed by its cause, D. Here, what we should always remember is that if the chief really has the causal belief if later D causes previous B, then he will take it for granted that the occurrence of previous B must be conditioned by the later occurrence of D; he will not be examining what will happen after previous B takes place, because if he was only trying to examine the relation between previous B and later D, then he can't be said to have an established causal belief of the kind later D causes previous B, as gained by his previous experience, and this is not the case of the chief as we

know it and as Dummett originally presented it, i.e., as a case of an already existing causal belief of the kind if later D causes previous B.

In short, the agent needn't really be ignorant of the occurrence of the past event to have a rationale for believing that there is no logical contradiction in believing that this event can be brought about by doing something in the present; this ignorance applies equally to all ordinary ways of being informed about the past, wither via memory or others' testimony. But more importantly, the agent who holds such a belief can't really be said to be ignorant of the occurrence of the past event, because the occurrence of the past event will fall under his knowledge in intention. The difference is only that this knowledge in intention is directed towards the past instead of being directed towards the future.

Dummett draws this conclusion in "Bringing about the past" by saying,

> My conclusion therefore is this. If anyone were to claim, of some type of action A, (i) that experience gave grounds for holding the performance of A as increasing the probability of the previous occurrence of a type of event E; and (ii) that experience gave no grounds for regarding A as an action which it was ever not in his power to perform, then we could either force him to abandon one or other of these beliefs, or else to abandon the belief (iii) that it was ever possible for him to have knowledge, independent of his intention to perform A or not, of whether an event E had occurred. (Dummett 1980: 349)

Let's read what Dummett had to say about that in his paper "Causal loops",

> Originally we made the natural assumption that we could, on occasions, know whether or not F had occurred independently of our intentions — precisely the assumption that cannot be made in relation to any future event which we believe ourselves to have the means of bringing about or of preventing. It was because we made that assumption that we were able to establish the correlation between the performance of B [present action] and the previous occurrence of F [past event]. It was also because we made this assumption that we took it for granted that there was no point in trying to bring it about that F occurred when we have clear evidence that it did not. Although we placed sufficient reliance on the correlation between B and F for the performance of B to count as increasing the probability that F occurred in cases in which we had no evidence of an ordinary kind about whether it did or not, we trusted such evidence so much more than we trusted the correlation

> that the performance of B did not significantly affect our estimate of the probability of F's having occurred in cases in which we possessed that evidence, even though we knew that evidence for a past event can sometimes prove mistaken. (Dummett 1996: 361)

For Dummett, the hindrance that prevents us from believing that the concept of affecting the past can be made sense of lies in our belief that knowledge in intention can be directed towards the future but can't be directed towards the past, due to the belief that knowing past events can have other more objective informing resources other than our knowledge in intention. But this is a mere psychological effect of having memory and not having foreknowledge, and in some special cases, where our experience repeatedly keeps telling us that performing certain type of actions in the present has been associated with the occurrence of certain types of events as previously observed to be taking place in the past, should be sufficient to make us take our knowledge in intention when directed to the past as being as reliable as our knowledge in intention as directed to the future.

One challenging question that one might be motivated to ask in response to my interpretation to Dummett's view on the ignorance condition is the following: should knowledge in intention be considered as knowledge? What I have asserted is that the agent who believes in the consistency of the idea of attempting to bring about a past type of event by doing something in the present can't be said to be ignorant of the past type of event, because that type of event falls under his knowledge in intention. The same is true for an agent who tries to cause something in the future by doing something in the present, his knowledge of that thing can be said to fall under his knowledge in intention. This means that, in both the cases of the future and of the past, the agent can't be said to be totally ignorant of the type of event he is trying to bring about. But what if knowledge in intention is not to be considered as knowledge? Here in my response I need to make it clear that, for Dummett, knowledge in intention is not to be considered less objective than knowledge as acquired by ordinary ways of being informed, and any reading that fails to appreciate this would end up misinterpreting Dummett on this point. However, the objectivity of knowledge in intention might be more easily defended in the case of the future, where this knowledge not only is acknowledged by Dummett, but also acknowledged

as a kind of knowledge that can contradict the other ordinary ways of knowing the future, such as prediction as based on causal laws. As Dummett makes clear, an agent who has the intention to bring about some effect in the future, that he believes to be able to bring about, can't fulfill the request to bring about that effect, when he knows by ways other than his intention that this effect will not take place, because his knowledge in intention and his knowledge in prediction will contradict each other, and he will eventually have to appeal to one of them. This appeal makes it apparent that knowledge in intention is not to be considered as less objective than knowledge as based on prediction in the case of the future. In "Bringing about the past" Dummett says,

> If someone believes that a certain kind of action is effective in bringing about a subsequent event, I may challenge him to try it out in all possible circumstances: but I cannot demand that he try it out on some occasion when the event is not going to take place, since he cannot identify any such occasion independently of his intention to perform the action [...] I cannot be asked to perform the action on some occasion when I believe that the event will not take place, when this knowledge lies in my intention to prevent it taking place; for as soon as I accede to the request, I thereby abandon my intention (Dummett 1980: 344)

This means that knowledge in intention plays a role that is no less objective than knowledge as based on prediction when both are applied to future circumstances, and of course the term 'objective' as applied to the way we know the future is relative since we don't have foreknowledge. Now, if knowledge in intention is applied to the past the way it is applied to the future, the result would be in favor of the argument for the consistency of backwards causation; because in some cases, as Dummett makes clear, the intention to cause the past event forms a ground for believing in the occurrence of that event, similar to what takes place regarding future events. Dummett says (for a previous event F and a later act B) "the intention to do B becomes itself a ground, in some cases, for supposing that F has occurred" (Dummett 1996: 369).

But this is not the way things tend to be, because while we separate our knowledge in intention from our knowledge of the occurrence of past events, we find out knowledge in intention hardly to be separable from our knowledge of the occurrence of future events.

This is because, as asserted before, we have memory but we don't have foreknowledge. But this doesn't mean that knowledge in intention is to be considered less objective than knowledge as based on prediction when both are applied to the future. Dummett says,

> The difference between past and future lies in this: That we think that, of any past event, it is in principle possible for me to know whether or not it took place independently of my present intentions; whereas, for many types of future event, we should admit that we are never going to be in a position to have such knowledge independently of our intentions (if we had foreknowledge, this might be different). (Dummett 1996: 349)

What is still important to be stressed here is to make a distinction between the natural attitude of the agent after *objectively* being informed by the ordinary ways of the occurrence or the nonoccurrence of past events, and what might be his rational attitude. Because, as Dummett makes clear, it is the natural response of us to lose our motivation to perform an action as designed to bring about previous types of effects after being informed of their nonoccurrence. But the non-naturalness of the response of a person who continues trusting his knowledge in intention to perform the action, more than trusting the ordinary objective way of being informed, doesn't imply the logical inconsistency of such a response. In "Bringing about the past", as we have read, Dummett describes the attitude of a person who will continue trusting his knowledge in intention more that the ordinary way of being informed as being "paradoxical and unnatural to us" even if he can't see any "rational considerations which force him out of this position" (Dummett 1980: 349).

In "Causal Loops" Dummett explains clearly the psychological reasons that prevent us from believing in the consistency of backwards causation. He clarifies the fact that these reasons are not based on logical or metaphysical grounds as he says,

> Thus what stands in the way of our supposing it rational to do anything in order that something else should previously have occurred is not the logical fact that the event in question has already either occurred or not occurred, or the metaphysical status of the past as fixed, in contrast to the fluid condition of the future, but our assumption that, of any past event, we may have evidence for its occurrence or nonoccurrence whose strength can be estimated independently of our intentions. This assumption is, of course, based, not only on causal connections from

> earlier to later, but on the absence of any comparable connections in the reverse direction — that is, connections that we might use to attempt to bring it about that certain events had previously occurred. Just because the assumption is deeply engrained in us, we should feel the strongest psychological resistance to recognizing any such connection; but, were we to recognize one; we should have, to that extent, to modify that assumption. This would profoundly alter our conception of evidence about the past, but it would not produce conceptual chaos (Dummett 1996: 363)

In making a distinction between causes that precede their effects and causes that follow their effects (quasi-causes), Dummett refers to the fact that quasi-causes appear redundant when we know that their effects have taken place. "We must compare with the effectiveness of quasi-causes the effectiveness of causes. A quasi-cause appears redundant when we know that the wished-for effect has taken place" (Dummett 1980: 331). The word 'appear' is not to be overlooked here because it is not the case that knowing the occurrence of past effects makes the quasi-causes redundant; it only makes the quasi-causes appear redundant.

Conclusion

In this paper I defended the argument for the consistency of believing in backwards causation against the objection from the ignorance condition. This objection, I argue, is based on a misinterpretation of Dummett's proposal. Therefore, in my defense, I presented an interpretation of Dummett's position in which I tried to clear up the confusion regarding the ignorance condition as a condition wrongly believed to be required by the agent to believe in the consistency of backwards causation. In my argument, I defended the view that, not only the agent doesn't need to be ignorant of the occurrence of the past event to have a rationale for believing in the logical consistency of attempting to bring about the occurrence of that past event by doing something in the present; in addition, such an agent can't be said to be really ignorant of that event, as long as bringing about that event lies in his knowledge in intention. I depended in my argument on the distinction that can be applied to both future and past events between two kinds of knowledge: knowledge in intention and

knowledge as acquired by ordinary ways of being informed.[3]

Abla Hasan
University of Nebraska-Lincoln
Office 1025 OLDH
Modern Languages and Literatures
1111 Oldfather Hall
660 N 12th St
Lincoln NE 68588-0315
abla.hasan@unl.edu

References

Dummett, Michael. 1980. Bringing About the Past. In *Truth and Other Enigmas*. Harvard University Press.

Dummett, Michael. 1980. Can an Effect Precede its Cause? In *Truth and Other Enigmas.* Harvard University Press.

Dummett, Michael. 1996. Causal Loops. In *The Seas of Language*. Oxford University Press.

Gorovitz, Samuel. 1964. Leaving the Past Alone. *The Philosophical Review* 73 (3): 360-371

Hanoch, Ben-Yami. 2007. The Impossibility of Backwards Causation. *The Philosophical Quarterly* 57 (228): 439-455.

Peijnenburg, Jeanne. 2006. Shaping your own Life. *Metaphilosophy* 37 (2): 240-253.

[3] I am grateful to Prof. Edward Becker, University of Nebraska-Lincoln, for his valuable discussions.

The Misuse and Failure of the Evolutionary Argument

Joseph Corabi
Saint Joseph's University

BIBLID [0873-626X (2014) 39; pp. 199-227]

Abstract
The evolutionary argument is an argument against epiphenomenalism, designed to show that some mind-body theory that allows for the efficacy of qualia is true. First developed by Herbert Spencer and William James, the argument has gone through numerous incarnations and it has been criticized in a number of different ways. Yet many have found the criticisms of the argument in the literature unconvincing. Bearing this in mind, I examine two primary issues: first, whether the alleged insights employed in traditional versions of the argument have been correctly and consistently applied, and second, whether the alleged insights can withstand critical scrutiny. With respect to the first issue, I conclude that the proponents of the argument have tended to grossly oversimplify the considerations involved, incorrectly supposing that the evolutionary argument is properly conceived as a non-specific argument for the disjunction of physicalism and interactionist dualism and against epiphenomenalism. With respect to the second issue, I offer a new criticism that decisively refutes all arguments along the lines of the one I present. Finally, I draw positive lessons about the use of empirical considerations in debates over the mind-body problem.

Keywords
Mind-body problem, epiphenomenalism, evolutionary argument, William James, physicalism

Introduction

The evolutionary argument purports to be an argument against epiphenomenalism — the thesis that mental states and events have no causal effects.[1] The argument claims that epiphenomenalism can be

[1] Later, I will also make clear that I assume that epiphenomenalism is commit-

Disputatio, Vol. VI, No. 39, November 2014
Received: 26/08/2013 Revised: 08/11/2013 Accepted: 30/05/2014

disconfirmed on empirical grounds, rather than merely being counterintuitive. The evolutionary argument has a distinguished history, being introduced by Herbert Spencer (1871), and having been defended — in one form or another — by William James (1890), Karl Popper (Eccles and Popper 1977), and others. Here is a passage from James's classic statement of the argument:

> There is… [a] set of facts which seem explicable on the supposition that consciousness has causal efficacy. *It is a well-known fact that pleasures are generally associated with beneficial, pains with detrimental, experiences.* All the fundamental vital processes illustrate this law. Starvation, suffocation, privation of food, drink and sleep, work when exhausted, burns, wounds, inflammation, the effects of poison, are as disagreeable as filling the hungry stomach, enjoying rest and sleep after fatigue, exercise after rest, and a sound skin and unbroken bones at all times, are pleasant. Mr. Spencer and others have suggested that these coincidences are due, not to any pre-established harmony, but to the mere action of natural selection which would certainly kill off in the long-run any breed of creatures to whom the fundamentally noxious experience seemed enjoyable. An animal that should take pleasure in a feeling of suffocation would, if that pleasure were efficacious enough to make him immerse his head in water, enjoy a longevity of four or five minutes. But if pleasures and pains have no efficacy, one does not see… why the most noxious acts, such as burning, might not give thrills of delight, and the most necessary ones, such as breathing, cause agony.[2]

While different figures have formulated the argument in subtly different ways, all of the ones following James's style have taken the central insight involved to be the basis for an argument for the causal efficacy of qualia; this central insight is that epiphenomenalism leaves the smooth correlation between negative qualia and harmful stimuli unexplained. Since all forms of interactionist dualism and virtually all forms of physicalism hold that qualia *are* causally efficacious, and all forms of epiphenomenalism hold that they are not, the argument is uniformly taken to be a non-specific argument for the disjunction of interactionist dualism and physicalism, and

ted to robust dualism of the sort proposed in Chalmers 1996. Serious defenders of epiphenomenalism have included Thomas Huxley (1874), Frank Jackson (1982), and William Robinson (2004).

[2] James (1890: 143-4), emphasis in original.

against epiphenomenalism.[3]

Although clever, the evolutionary argument has aroused its fair share of suspicion. The criticisms of it in the literature have been diverse, and also far from decisive. (See, for instance, Broad 1925, Jackson 1982, Van Rooijen 1987, Lindahl 1997, and Robinson 2003, 2007.)[4]

Given this background, my aim in this paper is twofold. First, I will show that the evidence the argument employs has been mishandled, even if we grant the important assumptions of the argument. Contrary to what its traditional proponents have led us to believe, it is not best conceived as a straightforward argument for the efficacy of qualia, and hence as a non-specific argument for the disjunction of interactionist dualism and physicalism. The matter is more subtle than this, and I will explain how the distinct kinds of evidence the argument employs pull us in a different direction from what someone like James supposed.

Second, once the traditional oversimplifications have been noted and an improved version formulated, I offer a new objection to the argument that decisively refutes it (or refutes it in anything like a traditional form, at least), by making clear once and for all the central mistake that plagues it. (The process of sorting out the earlier confusions will help to focus our efforts.) I will make the case that the central mistake lies in accepting one assumption in particular that is unjustified and almost certainly false. Although my primary aim

[3] I say that virtually all forms of physicalism hold this because there are a few physicalist views that hold that qualia are inefficacious because the neural states they supervene on (or are identical to) are physiologically cut off from the production of behavior. Such views are extremely rare and (relatedly) not usually considered plausible, so I ignore them here. Also, as I discuss below, for the purposes of this paper I do not classify as epiphenomenalist those physicalist theories that have trouble countenancing the causal efficacy of qualia for subtle metaphysical reasons (such as the ones that sometimes arise in connection with role functionalism) — I treat these as straightforwardly non-epiphenomenalist. Incidentally, I also assume throughout that all views must acknowledge the reality of qualia, even if they are ultimately reducible to or in some other metaphysically intimate way dependent on the physical. This is keeping with trends in the philosophy of mind over the past generation, where accounting for phenomenal consciousness has generally been considered of central importance.

[4] See also my response to Robinson (2007) in Corabi 2008.

here is critical, there are positive lessons to draw from this negative result; in particular, we gain some insight into what empirical considerations can genuinely help us to solve the mind-body problem.

I will begin by clarifying some important terms, and then formulate a canonical evolutionary argument against epiphenomenalism. I will then employ this canonical formulation to explore the ways in which the argument has mishandled the evidence, even conditional on the correctness of the assumption I will later subject to scrutiny. After providing an adjusted formulation that sidesteps these problems (though at the cost of complicating the commonly accepted conclusion of such arguments), I will then discuss the key assumption that drives these arguments toward their conclusion and explain why it drives them in this direction. The assumption is that physicalism, because it claims physical neural bases of qualia metaphysically necessitate the qualia themselves, thereby guarantees (for confirmation purposes) that all precise versions of physicalism will posit just this connection between the physical and the phenomenal.[5] Finally, I will explain why this assumption is unjustified, and explore what lessons can be learned as a result.

Some preliminary matters and the canonical formulation

The evolutionary argument is an inference to the best explanation, and consequently involves the evaluation of numerous different hypotheses. Before presenting the argument, it will be important to get a feel for the various hypotheses that might explain the evidence that needs explaining.

When examining the evidence the argument considers, we are trying to decide between three competing general theories on the mind-body problem — physicalism, (dualistic) interactionism, and (dualistic) epiphenomenalism. Interactionism and epiphenomenalism, as I will understand them, are robust dualist views, which deny the metaphysical supervenience of qualia on the physical.[6] Interac-

[5] This assumption has often been left implicit by defenders of the argument, but we will see below that it is required to get the argument off the ground.

[6] A prominent example of the kind of dualism that I am assuming these views are committed to is the one defended by Chalmers (1996).

tionism and epiphenomenalism differ from one another only in their views about the causal efficacy of qualia — interactionism accepts that at least some qualia are causally efficacious with respect to the physical, while epiphenomenalism denies that any qualia are causally efficacious with respect to the physical.[7] I will understand physicalism, on the other hand, as any view which accepts that qualia metaphysically supervene on the physical.[8]

I will offer a couple of brief remarks on these positions before continuing on to the argument itself. First, it should be noted that my understandings of interactionism and epiphenomenalism focus on the efficacy of qualia, not mental states generally. This is convenient for present purposes because traditional evolutionary arguments have primarily paid attention to correlations between dangerous distal stimuli and various simple, somatic experiences (as the James passage above illustrates). They have largely ignored stimuli that cause more nuanced and complicated mental states, involving emotions like fear and anger (and whatever propositional attitudes are associated with these emotions). In any case, though, insofar as emotions have a phenomenological element, they will fall under the auspices of these definitions. Second, on certain ways of classifying mind-body theories, some views I am classifying as physicalist count as dualist or epiphenomenalist. What views are these? I am thinking of various versions of property dualism that arise from concerns about multiple realizability and from related sympathy for role functionalism.[9] I classify these views as physicalist because they share what is, for the purposes of this argument, the most important feature in common with views that are straightforwardly physicalist and

[7] A more leisurely presentation of these views (and of the evolutionary argument itself) can be found in Corabi 2011. I avoid a leisurely treatment here because of spatial constraints, and because the treatment appears elsewhere.

[8] I will not attempt here to give any sort of precise characterization of the appropriate metaphysical supervenience relation. Such discussions are notoriously complicated and largely peripheral to present concerns. For a more detailed discussion, though, see Corabi 2011. As noted earlier, I also assume throughout that all physicalist views allow for qualia to play a causal role in behavior. (See my subsequent remark for how I am treating various forms of role functionalism and non-reductive physicalism.)

[9] For examples of such views, see Yablo 1992.

non-epiphenomenalist — they affirm the metaphysical necessitation of qualia by their physical neural bases and causation of physiological events in the nervous system (and ultimately behavior) by those physical neural bases. (It will become clear later why this shared feature is important.)

Now that we have seen the various general positions we are sorting through, let us examine the argument itself. Formulations of the argument have often been fairly quick and breezy, requiring the reader to fill in a number of important details and background assumptions. It will thus be useful to formulate a version from the ground up, making explicit as much detail as will be needed for our purposes, and so we will begin by looking at a formalized version of the traditional Jamesian version of the argument. (As mentioned previously, in the next section we will see how the formulation of this argument needs to be revised in light of problems unrelated to the key assumption — about the relationship between metaphysical necessitation and confirmation — but which are still of central concern.)

As noted above, the traditional evolutionary argument is essentially an abductive argument in favor of both physicalism and interactionism, and against epiphenomenalism. It attempts to show that physicalism and interactionism, because they allow for qualia to play a causal role in the physical world (including in behavior, presumably), lead us to expect the evidence we actually find, while epiphenomenalism does not. Hence, they are each confirmed and epiphenomenalism alone disconfirmed.

What is this evidence? It is of two kinds. The first is correlations between distal stimuli and qualia, and the second is what behaviors organisms display when exposed to various kinds of stimuli (and the grounding of those behaviors in the physiology of the nervous system).

In the process of formulating our canonical version of the argument, I will make use of two important principles. First, we should always use the most determinate evidence available. So, instead, of merely using evidence like 'sharp cuts to the arm result in avoidance behavior and are mediated by unpleasant qualia', we should use evidence like 'sharp cuts to the arm of determinate type *t* result in avoidance behavior of determinate type *b* and are mediated by qualia

of determinate type *q*.'[10] In addition, data about detailed physiological transitions in the nervous system of the organism should also be included (insofar as we know what they are). Second, I will view confirmation in an explicitly Bayesian fashion. Although there are competing theories of confirmation, Bayesianism has the advantage of allowing us to set up models that make it easier to visualize the confirmation process in action. Moreover, in the context of an argument like this one, the choice of a confirmation framework is unlikely to make any substantive difference, so presupposing a Bayesian framework will not involve smuggling in any controversial assumptions.

Now, the way the argument reaches its conclusion is to maintain that P(*e*/physicalism) and P(*e*/interactionism) are similar to one another and each is significantly greater than P(*e*/epiphenomenalism), where '*e*' denotes the relevant evidence about physiological transitions and correlations between qualia and distal stimuli.[11] It is a fundamental tenet of Bayesian confirmation theory that a piece of evidence confirms a hypothesis (i.e., makes it more likely to be true) if and only if the hypothesis is more likely on the evidence than the hypothesis is on the lack of the evidence. In turn, this relationship holds if and only if the evidence is more likely on the hypothesis than on the hypothesis's negation.[12] To put it more formally: $P(h/e) > P(h)$

[10] This is essentially because using less determinate evidence can lead to counter-intuitive confirmation results. It is true, of course, that we do not always use the most determinate evidence in our everyday abductive inferences, or even our scientific abductive inferences. But it will turn out that, in every context where we rely on less than fully determinate evidence, this is because there are either great practical difficulties in obtaining the fully determinate evidence, or else it is inconvenient to use such fully determinate evidence and it seems very unlikely that fully determinate evidence would lead to a different conclusion than the less determinate evidence we do use. I discuss these issues in more detail in Corabi 2011.

[11] For the sake of simplicity, I omit consideration of background knowledge here. I also intend the probabilities in question to be understood epistemically, as what are often called 'degrees of belief'. I will not attempt to tackle complicated probability issues here, however — I think the relevant notions are clear enough intuitively for the limited purposes of this paper.

[12] There are, of course, numerous qualifications to this thesis, but none of them is relevant for present purposes.

iff P(e/h) > P(e/~h), where e is the evidence and h is the hypothesis. But that is exactly what the above conditional probabilities are implying, of course — that the evidence is much more to be expected if one of physicalism or interactionism is true.[13]

As I mentioned earlier (and as the James passage indicated), the reason for drawing this conclusion is that when we examine the evidence, we are struck by two things. First, we are struck by the appropriateness of most of the behaviors we have when confronted by dangerous (and helpful) stimuli.[14] *Prima facie*, at least, this is not surprising on any of the hypotheses; after all, we would not be here if our ancestors had responded inappropriately to burns, cuts, and insect bites. But what is more interesting is the close correlation between dangerous stimuli and experiences that feel unpleasant in some hard to describe, but nevertheless very fundamental, sense. (These experiences are unpleasant not merely in the sense that they are not pleasant, but that they are positively "nasty" in their phenomenology.)

Here is where the traditional Jamesian argument gets its bite — if physicalism or interactionism were true, this "match" between qualia and stimulus would seem to be perfectly appropriate, since according to these views qualia exert a causal influence on behavior. Thus, if we (or our ancestors) felt something other than sharp pain when we were cut on the arm by a sharp knife, we would probably treat further cuts

---

[13] The argument claims that the evidence is much more likely conditional on interactionism, for instance, than on interactionism's negation, because interactionism's negation is the disjunction of physicalism and epiphenomenalism. Although the evidence would be likely conditional on physicalism, it would not on epiphenomenalism.

[14] For simplicity's sake, I will simply focus on the case of dangerous stimuli, though most of what is said can be applied straightforwardly mutatis mutandis to the case of helpful stimuli. Incidentally, there are cases where our dispositions are not so appropriate, of course. Take, for instance, many people's standing disposition to eat fatty foods when presented with them or to avoid vigorous exercise and painful immunizations. These cases are the rare exception rather than the rule, and most likely can be explained in a variety of ways. For instance, they may be explained by the fact that our ancestors lived in a different evolutionary environment than we do, that processes other than natural selection are at work in evolution, and that long-term individual survival is not always the goal of selection pressures. I will not speculate any further here, though, on how these explanatory stories might go.

too nonchalantly, or perhaps even seek them out, since whatever qualia we were experiencing would not motivate us to avoid the stimulus with sufficient urgency. Needless to say, this would quickly remove us from the gene pool! (We need not look to fanciful hypothetical examples to make this point. Although not precisely analogous, the tragic circumstances of many sufferers of congenital insensitivity to pain illustrate the dangers of being incapable of nociception.) Thus, if one of these hypotheses were correct, it would allegedly lead us to expect exactly what we find, which is what a high conditional probability of the evidence on the hypothesis indicates.

If epiphenomenalism were true, though, things would be different. Because epiphenomenalism entails that qualia have no causal influence on behavior, we get the intuition that qualia could be varied greatly without changing behavior at all. For example, an individual could easily feel ecstatic pleasure when cut by a knife, and still behave in exactly the same way as in the actual world. Thus, there would be no special reason to think the actual stimulus-phenomenology correlations would hold if epiphenomenalism were the case, hence the reason for the lower conditional probability of the evidence on the hypothesis.

A good (albeit idealized) way to think of the confirmation process is to envision each general hypothesis (e.g., epiphenomenalism) as a disjunction of highly determinate versions of that hypothesis, each of which specifies the history of the world in maximal detail.[15] Each of

[15] A reviewer objected that general hypotheses are not disjunctions of highly specific determinate hypotheses. Consider, for instance, the theory of plate tectonics. Surely it is ludicrous to suppose that the theory of plate tectonics is composed of a disjunction of a myriad of ultra-determinate theories specifying slightly different microscopic paths of plate movement. Worse, it seems preposterous to suppose that such theories would specify the entire history of the world in this level of detail! I respond by conceding that there is wisdom in this suggestion. For practical purposes, we do not specify theories at this level of detail because we do not have the time, memory, or computational capacity to concern ourselves with intricacies that will make no difference to our ability to assess general hypotheses. (This is because typically there will be no differences in what rival general hypotheses predict about events that are unrelated to the main phenomena they are designed to be theories about — a highly determinate version of plate tectonic theory can predict the movement of a specific atom in outer space just as easily as a highly determinate version of a rival "seafloor spreading" theory can, and

these determinate versions of the hypothesis will start off with an intrinsic probability, and the probability of the general hypothesis will be the sum of these smaller probabilities (since each of the determinate versions is mutually exclusive and together they exhaustively characterize the general hypothesis — if the general hypothesis is true, exactly one of the determinate versions will be true). As determinate versions of different hypotheses are ruled out by evidence that comes in, the probability that accrued to them initially will be reassigned to the remaining determinate options (regardless of what general hypotheses they are determinate versions of), maintaining their ratios to one another. So, for example, if a determinate version of epiphenomenalism with probability *x* is ruled out, that *x* will be distributed to all the remaining determinate options while maintaining their relative relationships. If there is a determinate version of physicalism, for instance, with probability *y* and a determinate version of interactionism with probability 2*y*, then the version of interactionism will inherit twice as much of the *x* as the determinate version of physicalism.[16]

---

vice-versa.) But it is important to realize that we are only making a concession to convenience when we omit detail in this way. An infinitely computationally powerful Bayesian demon with infinite memory and speed would not take such shortcuts. An indication that we are merely making a concession to convenience is that, when we are alerted to a potential difference between two versions of a general hypothesis that might lead to differences of prediction or to ontological differences in what is being posited, we have no difficulty recognizing that our old theory was ambiguous between them, and hence (in a sense) a disjunction of them. When assessing theoretical issues, sometimes it is illuminating to make all of this explicit and dispense with concessions to practicality. The present investigation is such an occasion, because dispensing with these concessions allows us to concentrate carefully on the characteristic ways that the respective general theories think about the production of behavior and its relationship to qualia. (The spirit of my remark here is similar to that in the note above on the precision of our formulation of the evidence.)

[16] A couple of remarks are in order. First, I assume that each general hypothesis is a disjunction of finitely many unique determinate versions (or at least countably many). If there is an infinity of determinate hypotheses comprising each general hypothesis (particularly an uncountable infinity), then this will introduce substantial mathematical complications that are well beyond the scope of the present paper, although I do not suspect that dealing with them would alter any of the substance of the arguments I give. Second, there are niceties that need

If we apply what I have said to the specific evidence at hand, the claim on the part of traditional defenders of evolutionary arguments is that, when we take the evidence into account, a large portion of the probability previously accruing to determinate versions of epiphenomenalism shifts to determinate versions of interactionism and physicalism (with no corresponding movement in the opposite direction). This is because a much larger proportion of these determinate versions of epiphenomenalism conflict with our evidence.

So, let us sum up our formulation of the traditional Jamesian argument. We can call the evidence we are considering here 'C' — it is roughly that humans have tended to behave appropriately in the light of numerous selection pressures (and individuals continue to behave appropriately in the light of familiar selection pressures), that there is a fairly smooth correlation between stimuli that enhance reproductive fitness and pleasure, and that there is also a fairly smooth correlation between stimuli that are detrimental to reproductive fitness and pain:

(1) A hypothesis *h* is confirmed iff P(*e*/*h*) > P(*e*/~*h*).[17]
(2) A hypothesis *h* is disconfirmed iff P(*e*/~*h*) > P(*e*/*h*).
(3) P(C/physicalism) > P(C/epiphenomenalism)
(4) P(C/interactionism) > P(C/epiphenomenalism)
(5) Physicalism, interactionism, and epiphenomenalism are mutually exclusive and jointly exhaustive.[18]

to be introduced to make the sort of process I describe here fully adequate and precise. None of these are relevant for present purposes, though, and so I omit them to avoid unnecessary technicality. For a bit more discussion of some of these issues, though, and a helpful visual aid, see Corabi 2011. See also Meacham 2008 for a similar visual aid.

[17] To keep things simple here, I omit reference to background knowledge.

[18] It should be noted that I consider the general hypotheses I have labeled 'physicalism' and 'dualism' to be agnostic on the question of panpsychism, and so all of the general hypothesis under consideration here are also agnostic on that question, since all are varieties of physicalism and dualism that take no explicit stands on panpsychism. (In using the terms in this way, I am following common usage in the literature in recent decades.) I will be setting aside panpsychist versions of the respective hypotheses, however, since dealing adequately with them lies beyond the scope of the limited goals of the present paper. It might

From (3), (4), and (5):

(6) P(C/physicalism v interactionism) > P(C/~[physicalism v interactionism])

And:

(7) P(C/~epiphenomenalism) > P(C/epiphenomenalism)

So, from (1) and (6):

(8) Physicalism v interactionism is confirmed.

And from (2) and (7):

(9) Epiphenomenalism is disconfirmed.

I offer a word on the interpretive justification for this formulation, since as I alluded to above James himself does not explicitly express his reasoning in a Bayesian fashion, but we are imposing the Bayesian formalism on his argument to ensure that it has adequate precision. (1) and (2) are background Bayesian assumptions discussed previously. (5) is undeniable, and although not made explicit by James, is a belief that it is fair to assume that he held. This leaves (3) and (4). On a first glance, someone might object that James never mentions physicalism, interactionism, or epiphenomenalism. How, then, could (3) and (4) be what he intended to express? It is important to note that, in the passage quoted above, James speaks of a "set of facts which seem explicable on the supposition that consciousness has causal efficacy" — a set of facts that includes the unpleasantness people feel in the presence of burns, wounds, and starvation. Invoking evolution, he suggests that "these coincidences are due, not to any

---

be objected that this represents an inappropriate assumption under the circumstances, since James's own all-things-considered view was panpsychist. While it is true that James was a panpsychist, his reasons for embracing panpsychism had no connection to the argument we are examining, and so his endorsement of that argument can be treated on its own terms independently of issues surrounding panpsychism.

pre-established harmony, but to the mere action of natural selection which would certainly kill off in the long-run any breed of creatures to whom the fundamentally noxious experience seemed enjoyable." He ends by claiming that "if pleasures and pains have no efficacy, one does not see… why the most noxious acts, such as burning, might not give thrills of delight..." It seems clear, then, that James is asserting that natural selection would kill off in the long run any organisms that sought out such "noxious acts". But according to him, only views that maintain that consciousness has causal efficacy can "explain" why organisms would avoid noxious stimuli. What does 'explain' mean in this context? The only plausible candidate is that it means *successfully predicts* — we can see this by contrasting his assessment of this view with his assessment of the "no efficacy" view. On the no efficacy view, "one does not see" why burning might not easily be correlated with very pleasant experiences — in other words, the no efficacy view makes no prediction about what sorts of qualia we would find paired with these stimuli. But the no efficacy view is epiphenomenalism, of course, as we have defined it. There is no single general view that holds that qualia are efficacious, however — there are really two views, one dualist and the other physicalist. These are the physicalism and interactionism of premises (3) and (4). Physicalism and interactionism, according to James, successfully predict the evidence, because they posit that people who survive the natural selection process will have unpleasant experiences in response to noxious stimuli, and hence avoid those stimuli, because this is what would have motivated their ancestors to avoid those stimuli and keep the species alive. Epiphenomenalism, on the other, does not predict the evidence, because according to epiphenomenalism human behavior throughout the evolutionary process would have been the same no matter what the qualia were; hence modern humans could just as easily feel delight at being burned as excruciating pain if epiphenomenalism were true. In Bayesian terms, this is tantamount to saying that the conditional probability of the evidence given physicalism and given interactionism is higher than it is given epiphenomenalism.[19]

Addressing the substance of the argument, (1) and (2) are (at least

[19] I am grateful to an anonymous reviewer for spurring me to discuss these interpretive issues in greater depth.

in outline) uncontroversial principles, and (5) is — as I already mentioned — undeniable. As I alluded to above, in the next section I will discuss a problem with (6) — and by extension with (3) and (4), which lead to (6) — that requires us to adjust the formulation of the argument and move its conclusion away from what proponents of evolutionary arguments have generally assumed is the sensible one to draw. Once we then have a finalized version in place, we will be in a position to appreciate the relevance of the key assumption, as well as the difficulties with that assumption that ultimately doom all arguments of this ilk.

The central traditional confusion and the key assumption

Before proceeding to the central traditional confusion about the argument, a remark about a more peripheral confusion is in order and will help to focus our attention more squarely on the heart of the matter. The reader may have noticed that, when a precise version of the "evolutionary argument" is formulated, the evidence having to do specifically with evolution is at best superfluous and at worst a serious distraction. This is so for two reasons, one fairly superficial and the other deeper and more far-reaching in its implications. First, the survival of presently living persons in the face of environmental challenges (and the characteristic qualia they receive as part of those challenges) gives us plenty of evidence in the spirit of the evolutionary evidence — it is probably true, after all, that most adults would not still be around if they felt pleasure at (and tended to seek out) burns, cuts, and insect stings. (As previously mentioned, sufferers from congenital insensitivity to pain, while not exactly analogous to individuals with "inverted pain spectra", do give us reason to suppose the fate of such people would not be promising. We do not really need evolutionary evidence to convince us of the problems with seeking out detrimental stimuli.) Second, when we imagine fully determinate versions of epiphenomenalism, physicalism, and interactionism (where the histories of the world are spelled out in full detail in these hypotheses), we see immediately that any possible physical history of events *outside the brains of humans* will be captured by an epiphenomenalist hypothesis, a physicalist hypothesis, and an interactionist hypothesis, and moreover each of these three maxi-

mally specific hypotheses will have roughly equal probability going in *ceteris paribus*. (This is because all of the views agree that qualia are uninvolved in events occurring outside the brain, and the only place the views disagree with one another is over the nature and/or causal role of qualia.) Thus, any evidence pertaining to matters outside the brain will be dealt with isomorphically by each of the three general views. (When determinate hypotheses are ruled out as a result of gathering evidence about behavior or evolution, the losses will be felt in equal proportion by all of the respective general hypotheses, and so will be returned to them in equal proportion.) It is only evidence about the brain itself (and about qualia) that have any chance of confirming or disconfirming any of the general views, since these are the only places the views will find themselves in serious tension with one another. For this reason, in subsequent discussion I will minimize my presentation of evidence having to do with evolution (and behavior), focusing instead on key evidence about the brain and about qualia. (I will still discuss the external stimuli that qualia are correlated with, however, as this will make it easier to see the significance of the information about qualia we have at our disposal).[20]

Now that we have seen where to focus our attention, it turns out that there are two reasons why the argument's traditional conclusion — that the argument is strictly an argument against epiphenomenalism — is unjustified, even granting the argument's key assumption (which will be discussed later).

The first reason is that the argument relies on evidence having to do with physiological transitions within the organism (in response to stimuli and resulting in behavior), especially evidence about physiological transitions within the brain. These transitions will either strongly favor physicalism and epiphenomenalism together, or else interactionism alone.[21] This is because we will ultimately wind up discovering either that they are in keeping with how physical entities outside the brain behave or that they are not. (In other words, we will wind up discovering that the behavior of the atoms and mol-

[20] For more detailed discussion of these points about the dispensability of evolutionary evidence, see Corabi 2011.

[21] This is evidence is closely related (but not equivalent) to evidence for the thesis of the causal closure of the physical.

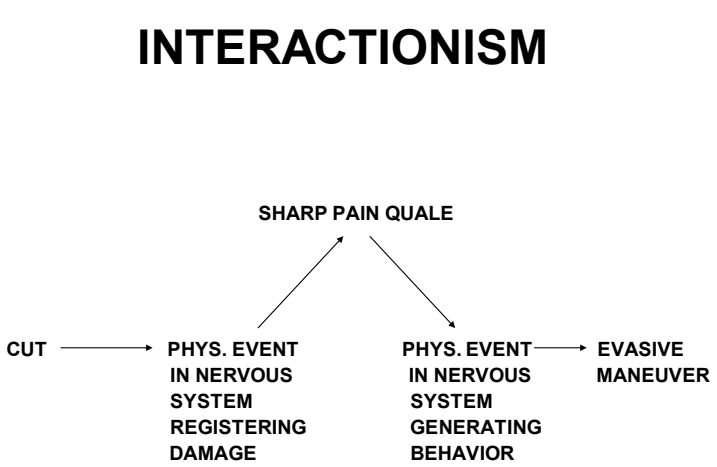
ecules inside the brain are just like the behavior of the atoms and molecules outside the brain under the same conditions, or else we will wind up discovering otherwise.) If they are in keeping with how these extra-cerebral physical entities behave, they will be the sorts of transitions physicalism and epiphenomenalism lead us to expect, since these views see behavior (and the physical events that lead to behavior) as ultimately governed by physical law alone. If they are not in keeping with the behavior of extra-cerebral physical entities, however, then this will strongly favor interactionism, since only interactionism leaves reasonable room for physical entities inside the brain to behave in a different fashion from those outside it. (This is because they are being "pushed around" by non-physical entities — namely qualia.)

The second, and for later purposes more important, reason why the argument's traditional conclusion is unjustified is that the argument relies on evidence having to do with the correlations between qualia types and distal stimuli. It turns out that when we consider the matter carefully, interactionism is subject to the same kinds of issues as epiphenomenalism where qualia "mixing and matching" is concerned — i.e., just as we have no special reason to expect unpleasant qualia to be associated with dangerous stimuli if epiphenomenalism is true, we have no special reason to expect it if interactionism is true either. This is roughly because interactionism (in most forms) posits two sets of contingent fundamental causal laws of nature where consciousness is concerned — a set of laws from physical to phenomenal (similar to epiphenomenalism), and then one from phenomenal back to physical.[22] Since they are metaphysically contingent, there appears to be no reason why these laws could not be varied to work harmoniously to produce adaptive behaviors in response to dangerous stimuli, and simply have the survival-conducive transitions causally mediated by different qualia.[23] So, for instance, if interactionism is

[22] Some forms of interactionism do posit non-mechanistic roles for qualia or other mental entities (such as, e.g., with robust agent causation views). In any case, I will set these aside for present purposes, mostly because dealing with them in full generality would take us far afield. I doubt, though, that anything about them would have a substantial impact on the basic force of my arguments.

[23] I assume throughout that fundamental laws of nature are metaphysically

true, then the actual process works like this when an organism is cut in the arm by a knife (where arrows indicate causal processes):

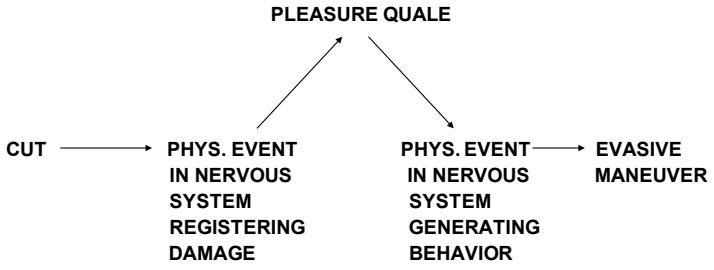


But it could have instead looked like this:

---

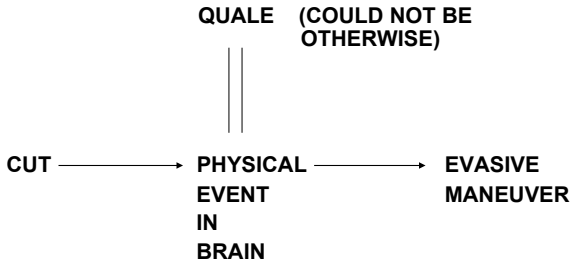
contingent, contra Shoemaker (1980), as well as what is often described as "realist" or (more informally) "oomphy" (or at least as describing oomphy causal processes). The argument can be paraphrased into a framework where the laws are treated as metaphysically necessary (so long as what properties are instantiated is not also metaphysically necessary) and perhaps also where the laws are Humean, but I will not speculate on the details of these paraphrases here. (The general idea of the necessitarian paraphrase is that there will be qualia properties that have identical "feels" to the actual ones, but which differ in their causal/nomic profiles. Thus, there will be possible worlds where such properties are instantiated, and these possible worlds will parallel the ones non-necessitarians believe in. In a standard non-necessitarian framework — where the causal/nomic profile of a property can vary from world to world — the same property would appear in many worlds, and would have many different causal/nomic profiles. In the necessitarian framework, it would be a different property in each of these possible worlds, but the centrally important feature would be preserved: the same range of causal/nomic roles matched up with the same qualitative feels.)

INTERACTIONISM ALTERNATIVE



So it is only physicalism that, by the argument's own lights, would lead us to expect the qualia/stimulus correlations we find, because only physicalism denies the metaphysical contingency between the neural base of a quale and the quale itself, as follows:

PHYSICALISM



So if the basic dialectical assumptions of the argument are correct, only physicalism will benefit from the qualia/stimulus correlation

evidence, and either physicalism and epiphenomenalism together or interactionism alone will benefit from the physiological transition evidence. But either way, it is very hard to see how physicalism and interactionism could be confirmed together and epiphenomenalism disconfirmed by itself, as the argument has typically concluded (represented in essence by (6) in the formulation of the argument in the last section). (The only way would be if physicalism benefited from the qualia/stimulus correlation evidence and interactionism benefited from the physiological transition evidence to just the right degree. But this possibility is so far-fetched, I will not even worry about it here.)[24]

So, to get a more precise feel for the relevance of these considerations, I will summarize an updated version of the Jamesian argument. Unfortunately, although many people have strong hunches about how the physiological transition evidence will turn out, at this point we have little information about the brain at a high enough resolution of detail to count as genuine evidence that can help in settling the question of whether physiological transitions will turn out to be those predicted by physicalism/epiphenomenalism or those predicted by interactionism. (In any case, as we have seen, assessing the impact of such evidence — if it does exist — is relatively straightforward.) Consequently, qualia/stimulus correlation evidence will

[24] I should briefly address an objection that may have popped up in the minds of some readers — what justifies us in supposing that if dualism is true (in either an interactionist or epiphenomenalist form), the neural base that actually generates (e.g.) a certain kind of pain could have just as easily generated a pleasure instead? I have two responses. First, there seems to be no obvious reason why not. Surely, there are some phenomenologies that would be either impossible for that neural base to generate, or at least intrinsically very unlikely — such as a complex visual phenomenology, for instance. This is because such a phenomenology would seem to require a different sort of information, or at least a great deal more information, than the neural base could reasonably encode. But why suppose pleasures and pains would be different from one another in this way? Second, and more importantly, such a variation in valence is not really required. Since it is physicalism's claim of metaphysical necessitation of the actual qualia by the actual neural bases that is doing the work, it does not ultimately matter what the character of these qualia is. All that matters is that there be a range of variations which are metaphysically possible if dualism is true, and no one would doubt that dualism allows for some variation, even if not as dramatic as flip-flops in valence.

be our main focus the rest of the way, and so I will formulate the argument so that it too focuses solely on these qualia issues. To keep things manageable, I will just suppose that the physiological transition evidence is totally up in the air, and so information about it cannot be taken into account at this stage.[25] Here is the argument, where 'Q' stands for the relevant qualia/stimulus correlation evidence:

(A) A hypothesis *h* is confirmed iff P(*e*/*h*) > P(*e*/~*h*).
(B) In general, a hypothesis *h* is disconfirmed iff P(*e*/~*h*) > P(*e*/*h*).
(C) P(Q/physicalism) > P(Q/epiphenomenalism v interactionism)
(D) Physicalism, interactionism, and epiphenomenalism are mutually exclusive and jointly exhaustive.

So, from (C) and (D):

(E) P(Q/~[epiph. v interactionism]) > P(Q/epiph. v interactionism)

And from (A), (D), and (E):

(F) Physicalism is confirmed.

And from (B) and (E):

(G) Epiphenomenalism v interactionism (i.e., dualism) is disconfirmed.

Although the above realizations damage the rationale for the standard conclusion of evolutionary arguments (i.e., that only epiphenomenalism is disconfirmed, not the disjunction of epiphenomenalism and interactionism), they leave the basic dialectical strategy essentially untouched in its core respects. Although the strategy

[25] To be perfectly satisfactory, this idea of being "totally up in the air" would have to be made more precise, but what we have should be good enough for present purposes.

does not support exactly the conclusion it was traditionally thought to, at this stage it nevertheless remains standing as a viable basis for an empirical argument designed to settle debate on the mind-body problem. Hence, from now on, I will focus on versions of the evolutionary argument that do not make the mistakes just discussed. Although these will differ from traditional versions of the argument in what mind-body theory/theories they conclude are confirmed by the actual findings (most likely, they will claim that only physicalism is confirmed), they will share with traditional versions the emphasis on the possibility of qualia "mixing and matching" to drive them to their conclusions.

Predictably, then, the qualia/stimulus correlations will be the crucial evidence in our subsequent discussion (i.e., the considerations that allegedly support (C) in the above argument). It is of paramount importance for the success of the argument that, because physicalism posits a metaphysical necessitation relation from physical neural base to quale and the alternatives do not, physicalism gains a decisive confirmation advantage where the qualia/stimulus evidence is concerned. Unfortunately, I will ultimately conclude that this crucial assumption cannot be successfully defended, and so the argument falls with it. Let us now turn to a more detailed examination of the assumption, and its bearing on the evolutionary argument.

The problem with the key assumption

The central point to note is that there is no issue about the conceptual or epistemic separability of qualia and physical events, even if physicalism is true. It is plainly apparent that even if physicalism is true, it is nonetheless conceivable in some sense that the physical neural base of an actual quale be associated with some other quale or no quale at all.[26] (To put things another way: we can imagine having *discovered* that the actual neural base of a certain kind of sharp

[26] I assume here that physicalism is a priori possible. If physicalism is demonstrably false a priori (as proponents of Knowledge, Zombie, and Structural Arguments have contended), then these evolutionary arguments will be unsuccessful anyway. (There may be lessons in the offing even for those who are persuaded of the truth of dualism a priori, but I will not speculate here. A bit of what I say in the conclusion addresses this issue.)

pain was actually associated with a pleasure, or with only dreamless sleep. The fact that we can imagine things having turned out this way indicates that the scenario in question is epistemically possible.) The only issue is whether or not this has an impact on confirmation, and makes us judge the conditional probability of the evidence on physicalism (significantly) lower as a result. Thus far, we have been following the argument above in supposing that this is not so — that metaphysical necessity is also necessity for confirmation purposes.

Reflecting a bit more on the situation, though, there does not seem to be any particularly good reason to doubt that epistemic contingency rather than metaphysical contingency should be the relevant modality where confirmation is concerned — after all, confirmation is an epistemic matter *par excellence*. Since it seems that alternate determinate physicalist hypotheses are nevertheless epistemic possibilities, ruling them out should have an adverse effect *ceteris paribus* on the likelihood of physicalism being true. This has the implication, though, that wherever some determinate version of epiphenomenalism or interactionism posits a correlation between a quale and an underlying physical brain event, there will be a parallel determinate version of physicalism that posits the same connection. This will ruin our justification for (C) in the argument of the previous section, because the only reason we had for thinking P(Q/physicalism) was greater than P(Q/epiphenomenalism v interactionism) in the first place was that these dualist hypotheses allowed for a different quale to be associated with the same underlying physical brain state (and ultimately the same external stimulus and behavior), while physicalism allowed for only the actual quale to be associated with it. This meant that, when the real quale was observed and its association with that physical brain state noted, many previous determinate epiphenomenalist and interactionist options were ruled out, but no physicalist ones were. But now, given that we recognize physicalist options corresponding to these epiphenomenalist and interactionist ones, there is a parallel process across the board, and no general hypothesis gains or loses any ground.

Objections

Before summing up the findings of the paper and discussing broader

lessons, I will deal with some objections and big picture challenges:

(1) *In spite of what you say, metaphysical possibility is really what is relevant to confirmation, not epistemic possibility.*

Response — The best way to answer this objection is to point out that it would have terribly counterintuitive consequences. To see this, consider how this approach would work in a field far-removed from the mind-body problem:

Everyone believes the identity claim 'water = H_2O' has been highly confirmed. And presumably the reason it has been highly confirmed is that it began with a certain intrinsic probability, and then as evidence was gathered and alternative identity claims were ruled out (such as, for example, 'water = XYZ'), it inherited probability from these ruled out claims via the process previously discussed. But if the proposal on the table is correct, then this cannot be the right diagnosis, since no coherently thinking agent would recognize the metaphysical possibility of all the competing identity statements at once, since each is a metaphysically necessary truth if a truth at all, and the truth of each one is incompatible with the truth of the others. The allegedly correct diagnosis is rather that a more general claim, something like 'water is identical to a physical substance', was confirmed because its intrinsic probability was maintained as the evidence was taken into account while the intrinsic probability of other options ('water is an optical illusion' (e.g.); 'water is a chemical mixture') was siphoned off. All the while, potential determinate versions of 'water is identical to a physical substance' were being narrowed down, till only the one remained.

Convoluted as this account is, it gets even worse when we contemplate the confirmation of the specific proposition 'water = H_2O'. Although the convoluted account at least produces the right answer to the question 'was the proposition "water is identical to a physical substance" confirmed?' (i.e., yes!), it cannot produce the right answer to the question of whether 'water = H_2O' was confirmed. Rather than giving the obviously correct answer that everyone agrees on — i.e., that the proposition was confirmed — it must claim that 'water = H_2O' had no intrinsic probability, and only can be said to have a probability at all when it is the only determinate option left standing among the versions of 'water is identical to a physical substance.'

(Recall that confirmation is essentially a raising of the probability of a hypothesis by considering the evidence. But if 'water=H_2O' had no probability along the way, then there was no probability to raise.) The ridiculousness of this conclusion is too much to stomach.

(2) It seems like the overarching complaint behind the evolutionary argument is simply the all-too-familiar explanatory gap, because all that is ultimately at issue is the relationship between physical brain states and qualia. So then why is it even worth talking about?[27]

Response — It is true that issues surrounding the epistemic separability of qualia and physical brain states wind up being of crucial importance to the evolutionary argument.[28] (This is because the argument ultimately relies on there being a crucial disanalogy between physicalism and dualism — namely, that dualism leaves it metaphysically open what qualia will be instantiated when a particular physical brain profile is instantiated, while physicalism does not.) However, there are numerous reasons the argument is worth discussing in spite of this fact. First, historically no one seems to have noticed the crucial role of explanatory gap considerations in it. Seeing that the argument has been influential (defended, in one form or another, by several luminaries), it seems worth the trouble of clarifying its relationship to other issues that are relevant to the mind-body problem. Second, although it may exploit explanatory gap considerations, those considerations are used in a very different way in the evolutionary argument than they typically are. Normally, the epistemic separability of qualia from the physical is used as the basis for a pro-dualist argument, whereas with the evolutionary argument it is used as the basis for an argument for the causal efficacy of qualia (whether those qualia are ultimately construable physicalistically or not), a largely separate matter. We even saw that once other con-

[27] I am grateful to an anonymous reviewer for spurring me to clarify the discussion of this objection.

[28] For this reason, my discussion is not meant to apply to those who claim that we can infer the presence of the relevant conscious states a priori from physical descriptions. Such physicalists are rare nowadays and I believe their position is implausible, although I readily admit that it is difficult to give convincing arguments against it (largely owing to the fact that crucial premises in any such argument would be less secure than the conclusion itself).

fusions have been unmasked and peripheral issues set aside, in this context it really forms the basis of an anti-dualist argument (albeit an ultimately unsuccessful one)!

(3) Suppose that we will eventually discover that the physical brain state underlying a particular kind of negative qualia (QN) is physical neural base 1 (NB1), and the physical brain state underlying a particular kind of positive qualia (QP) is physical neural base 2 (NB2). The negative qualia are produced by a damaging stimulus (SD) and the positive qualia by a beneficial stimulus (SB). Consider what this will do to the confirmation of the various general hypotheses. Physicalism has only two possible determinate versions to start with — version A has (SD, QN, NB1) and (SB, QP, NB2), while version B has (SD, QN, NB2) and (SB, QP, NB1). (The difference is that version B has swapped qualia valences from version A.) Epiphenomenalism, on the other hand, has four possible determinate versions to start with — the parallels of A and B and also C (SD, QP, NB1) along with (SB, QN, NB2) and D (SD, QP, NB2) along with (SB, QN, NB1). (Epiphenomenalism also allows for the swapping of what physical brain states are correlated with what external stimuli, which is what gives us the two additional options.) But then physicalism and epiphenomenalism are not parallel after all — since the probability associated with physicalism is only split 2 ways initially and the probability associated with epiphenomenalism split 4 ways, physicalism receives confirmation and epiphenomenalism disconfirmation after all once we get the final evidence, since we rule out more determinate versions of epiphenomenalism than we do physicalism.[29]

Response — A first point to make about this objection is that its presentation of the evidence (and the determinate hypotheses on the table) is incomplete; for completeness (even setting aside physiological transition evidence), we would need not just the general positive/negative valence of the qualia, but much more detailed information about their nature (and the same goes for the stimuli).

However, we do not need to dwell on these issues to see the difficulty for this objection. The main problem is that there is no reason to believe in the sort of asymmetry the objection presupposes. Why, after all, would epiphenomenalism allow *a priori* for versions that allowed for mismatches between stimulus and physical

[29] Thanks to an anonymous reviewer for suggesting a discussion of this objection.

brain state (mismatches in the sense that these physical brain states would not lead to behavior conducive to reproductive fitness) that were not allowed by physicalism? Since the neural processing of data from stimuli and the subsequent behavior generated in response to those stimuli are both a matter of the activity of physical entities governed purely by physical laws according to both epiphenomenalism and physicalism, any possible physical arrangement of the brain and nervous system in response to stimuli will be represented by a possible determinate version of epiphenomenalism and a possible determinate version of physicalism.

The upshot — evolutionary arguments fail

As I have alluded to throughout the second half of the paper, the conclusion about the space of confirmation being the space of epistemic possibility is of crucial importance. It spells doom for the evolutionary argument. Although the inferences may already be clear, it is worth spelling them out explicitly.

Essentially, what the result we have arrived at does is strip physicalism of its ability to take advantage of the metaphysical contingency of the correlation between qualia and physical events on epiphenomenalism and interactionism. The metaphysical contingency of these correlations on these views, and the metaphysical necessity of them on physicalism, is of no significance to the argument. Because the correlations are epistemically contingent on all the views, and because the space of confirmation is the space of epistemic possibility (as we saw above), any time observation rules out an epistemically possible correlation, there will be analogous "loss" by all the general hypotheses, and thus they will all remain equal to where they were beforehand. Granted, there may be room for subtle differences between the hypotheses (in particular, between interactionism and the other options, owing to interactionism's added laws), but if they exist, these differences will be very subtle indeed, hardly enough to confidently ground any sort of argument against any of the views. To directly relate these considerations back to our updated Jamesian argument, its crucial premise — i.e., (C), that P(Q/physicalism) > P(Q/epiphenomenalism v interactionism) — is false. We have no reason to believe the evidence is more likely given physicalism than

it is given dualism.

Conclusion

At this point, we can stop and appreciate the positive lessons that can be salvaged from the demise of the evolutionary argument. Appreciating the flaws of the argument can help us to clarify exactly what considerations are potentially fruitful in helping us to solve the mind-body problem and gain a better understanding of mental causation. While perhaps not essential, purifying the discussion in this manner can help to prevent confusion and distraction in future debates.

It appears that the only directly useful empirical considerations will be ones having to do with whether physical entities inside the brain consistently behave in the same ways as those outside the brain. If they do not behave in the same ways, then for the reasons outlined above, we will have considerable evidence in favor of interactionism. And alternatively, if they do, then we will have considerable evidence against interactionism, and hence in favor of the disjunction of physicalism and epiphenomenalism.[30]

The only other tools at our disposal for dealing directly with the mind-body problem are bread and butter *a priori* considerations.[31] Surely if physicalism is ruled out or made less palatable *a priori*, this will have significant effects on the intrinsic probabilities of the determinate physicalist options, and also on the intrinsic probabilities of

[30] There is another type of evidence that could potentially play a role. If it were found that distinctive (and fairly natural, joint-carving) qualia types did not correlate smoothly with any neural base types, this would be evidence for dualism over physicalism. This is because only dualism allows for the possibility of this sort of variation, though only intrinsically far-fetched versions of dualism predict this. In any case, virtually every indication we have suggests that we will not find this, and almost no one (physicalist or dualist) suggests otherwise, so I will not bother to consider the possibility further. (Note that the correlation in question here need only be one directional — 'if neural base n, then qualia q'. The converse sort of correlation could be ruined by multiple realizability, but this would not have an impact on the issues at hand.)

[31] I am counting as a priori here more than just inferential relations between concepts and the like. I am also including arguments and intuitions about the limitations of (e.g.) conceivability as a guide to possibility, and information about the broad nature of the physical.

the determinate dualistic hypotheses, since probability in this setting is a zero sum game (because the general theories are together mutually exclusive and jointly exhaustive).

In any case, all other matters aside, it is clear that empirical considerations of the sort adduced by James and other traditional proponents of evolutionary arguments against epiphenomenalism will not bear fruit. Those philosophers hopeful that the empirical evidence adduced in those arguments would shed light on these issues in the philosophy of mind will either have to go back to the drawing board, or return to old fashioned armchair philosophical theorizing.[32]

Joseph Corabi
Saint Joseph's University
Department of Philosophy
5600 City Ave.
Philadelphia. PA 19131 USA
jcorabi@sju.edu

References

Broad, C.D. 1925. *The Mind and its Place in Nature*. London: Routledge and Kegan Paul.

Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Corabi, Joseph. 2008. Pleasure's Role in Evolution: A Response to Robinson. *Journal of Consciousness Studies* 15: 78-86.

Corabi, Joseph. 2011. Why the Evolutionary Argument isn't really an Evolutionary Argument after all. *Journal of Consciousness Studies* 18: 44-65.

Eccles, J. and Popper, K. 1977. *The Self and its Brain: An Argument for Interactionism*. Berlin: Springer-Verlag.

Howson, C. and Urbach, P. 1996. *Scientific Reasoning: The Bayesian Approach*, *2nd edition.* Chicago: Open Court.

Huxley, Thomas. 1874. On the Hypothesis that Animals are Automata. Reprinted in *Collected Essays*. London, 1893-94.

Jackson, Frank. 1982. Epiphenomenal Qualia. *Philosophical Quarterly* 32:127-136.

James, William. 1890. *The Principles of Psychology*. Cambridge, MA: Harvard University Press.

Lindahl, B. I. B. 1997. Consciousness and Biological Evolution. *The Journal of Theoretical Biology* 187: 613-629.

[32] I am grateful to Brian McLaughlin, Susan Schneider, Audre Brokes, Jamie Hebbeler, and Todd Moody for helpful discussion and comments on previous drafts.

Meacham, C. (2008) Sleeping Beauty and the Dynamics of *De Se* Belief. *Philosophical Studies* 138: 245-269.
Robinson, William. 2003. Epiphenomenalism. In *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/epiphenomenalism/#Natural
Robinson, William. 2004. *Understanding Phenomenal Consciousness*. Cambridge: Cambridge University Press.
Robinson, William. 2007. Evolution and Epiphenomenalism. *Journal of Consciousness Studies* 14: 27-42.
Shoemaker, Sydney. 1980. Causality and Properties. In *Time and Cause*. Edited by Peter Van Inwagen. Dordrecht: Reidel.
Spencer, Herbert. 1871. *Principles of Psychology*. New York.
Van Rooijen, Jeroen. 1987. Interactionism and Evolution: A Critique of Popper. *British Journal for the Philosophy of Science* 38: 87-92.
Yablo, Stephen. 1992. Mental Causation. *Philosophical Review* 101: 245-280.

James's Evolutionary Argument

William S. Robinson
Iowa State University

BIBLID [0873-626X (2014) 39; pp. 229-237]

Abstract
This paper is a commentary on Joseph Corabi's "The Misuse and Failure of the Evolutionary Argument", this Journal, vol. VI, No. 39; pp. 199-227. It defends William James's formulation of the evolutionary argument against charges such as mishandling of evidence. Although there are ways of attacking James's argument, it remains formidable, and Corabi's suggested revision is not an improvement on James's statement of it.

Keywords
Epiphenomenalism, IBE, physicalism, pleasure, William James

In "The Misuse and Failure of the Evolutionary Argument", Professor Corabi offers a critique of James's argument against automatism (what today we call "epiphenomenalism"), and develops what he claims to be an improved version of the argument, even though, in the end, he believes it is a failure. In this response, I will try to bring out the strength of James's argument, and explain why I think Corabi's version is not an improvement.

I had better say at the outset that I am not persuaded that rejection of epiphenomenalism is forced upon us by James's evolutionary argument (hereafter, the EA). I have explained why in Robinson 2007 and shall not repeat that material here.[1] However, I do think that the EA is a formidable argument, and that James expressed it quite well. In the first part of these remarks, I will recast James's argument in a

[1] Professor Corabi has replied to my 2007 article in Corabi 2008. I have not published a response to this paper, and I will not comment on it here, except to say that I stand by my 2007 paper.

Disputatio, Vol. VI, No. 39, November 2014
Received: 10/06/2014 Revised: 22/09/2014 Accepted: 23/09/2014

style that will ease our discussion. (Readers may compare my version with James's paragraph quoted at the beginning of Corabi's paper.) Later sections will comment on Corabi's view of the EA.

1 James's argument

In what follows I will use 'approach' (to a stimulus) to include not only reduction of spatial distance, but also prolongation or repetition of proximity to a stimulus. Similarly, tendencies to 'avoid' will include tendencies to increase spatial distance, and tendencies to shorten or avoid repetition of proximity to a stimulus.

'Smooth correlation' (between utility and hedonic valence) will refer to the following facts. (a) For the most part, approaching stimuli that are beneficial to us (that enhance our fitness) is pleasant. For example, eating when hungry, and drinking water when thirsty, are pleasant. (b) For the most part, approaching stimuli that are detrimental to us is unpleasant. James cites consumption of alcohol to the point of drunkenness as an exception, but holds that such exceptions "relate to experiences that are either not vital or not universal" (James 1890, v. 1: 144). Neither Corabi nor I make any point that hangs on the exceptional cases.

There is, of course, also a high correlation between utility and behavior. For the most part, we do not approach the detrimental, and we do not avoid the beneficial. The heart of James's argument, however, turns on pleasure and pain, and I will use 'smooth correlation' for the match between the (positive or negative) utility of stimuli and the pleasure or pain that we get from their presence. I will reserve 'high correlation' for the utility/behavior correlation.[2]

With these understandings in place, we can set out James's argument as follows.

J1 There is a smooth correlation between utility of stimuli and their hedonic valence.

J2 If the pleasantness and painfulness of stimuli had no effects

[2] In the chapter that contains the EA, and in the first sentence of the key paragraph, James writes of "consciousness". However, the argument itself is entirely about pleasures and pains. My formulation of the argument reflects this fact.

in behavior, then there would be no explanation for the fact in J1.

J3 If the pleasantness and painfulness of stimuli do have effects in behavior, we can explain the fact in J1.

The envisaged explanation invokes natural selection. Organisms lacking the smooth correlation would soon die out, if pleasure and pain are efficacious, because they would approach the detrimental and/or avoid the beneficial. Organisms that did have the smooth correlation would be more likely to live long enough to pass their traits — including whatever grounds the smooth correlation — on to the next generation. On the other hand, if pleasure and pain are not efficacious, "one does not see" why there should not be serious mismatches between utility and hedonic value, and it would require an unscientific embrace of "*a priori* rational harmony" to explain the fact in J1.

James takes the implication of these premises to be obvious, but let us state the conclusion explicitly.

J4 The hypothesis that pleasures and pains are efficacious is a better hypothesis than the hypothesis that they are inefficacious.

James does not mention physicalism or interactionism, but both of these views accept efficacy of pleasures and pains. So, in the contemporary climate of opinion, it is reasonable to extend James's own conclusion in the following way.

(J5) Physicalism is a better hypothesis than epiphenomenalism.

(J6) Interactionism is a better hypothesis than epiphenomenalism.

And it is a consequence of these that

(J7) [Physicalism v Interactionism] is a better hypothesis than epiphenomenalism.

However, the key point that underlies (J5) through (J7) is the acceptance of efficacy for pleasures and pains; and this is clearly stated in J4.

It is well known that physicalism and interactionism have difficulties of their own. So, it may well be that (J5) and (J6) are false — false because, for reasons not mentioned in James's argument, they

turn out to be either worse hypotheses than epiphenomenalism, or equally problematic hypotheses, all things considered.

Plausibly, it is worries of this kind that lead Corabi to a desire to give an improved version of the EA. His discussion of interactionism can, I believe, be summed up as the point that natural selection will not explain why the laws from pleasure to approach, and pain to avoidance, are what they are. If laws are contingent, then there are alternative possible scenarios that will include the high correlation of utility and behavior, but lack the smooth correlation of utility with hedonic valence — namely, scenarios in which both the valence and the laws connecting valence to behavior are inverted. This point seems to me to be correct.

Corabi's discussion of physicalism is harder to summarize. I will return to it briefly after we have seen more of Corabi's framework. The point to notice for now is just that since (J5) and (J6) have problems of their own, it cannot be clear that James's argument to the best (or, at least, the better) explanation is altogether decisive. With regard to this point, Corabi and I are in agreement.

Our fundamental disagreement concerns the following claims made in the introductory section of Corabi's paper.

C1 James's argument "mishandles" the evidence.

C2 Corabi's "canonical" version of the argument is an improved version of the EA.

C3 The central mistake of both versions is accepting an assumption that is unjustified and almost certainly false. This assumption is that "all precise versions of physicalism will posit just this [i.e., the actual] connection between the physical and the phenomenal".

2 Difficulties about "the evidence"

In my view, trouble begins early in the section on "Some preliminary matters and the canonical formulation", when Corabi asks "What is this evidence [on which the evolutionary argument is based]?" Corabi's answer is that it consists of correlations between distal stimuli

and qualia, and between exposure to stimuli and behavior. Corabi then adopts the principle that we should always use the most determinate evidence available. Thus, we should avoid formulations in terms of sharp cuts, avoidance behavior and unpleasant qualia, and instead "use evidence like 'sharp cuts to the arm of determinate type *t* result in avoidance behavior of determinate type *b* and are mediated by qualia of determinate type *q*'."

This account of the "evidence" used in the EA seems to me to be profoundly mistaken. The evidence to which the EA appeals is the smooth correlation — the fact stated in J1. The gist of the argument is simply that efficacy for pleasures and pains allows for a better explanation of J1 than would any theory that denies such efficacy.

Let us introduce 'determinate stimuli-behavior-qualia statements' (hereafter, 'determinate SBQ statements') to refer to statements of the form illustrated by Corabi's example quoted a few lines above. What might determinate SBQ statements be evidence *for*?

It might be suggested that they are evidence for J1. This cannot be right. The smooth correlation asserted by J1 holds between utilities and hedonic valences. But determinate SBQ statements do not classify stimuli as beneficial or detrimental, and they do not classify qualia as pleasant or unpleasant.[3] Without such classifications, they are *irrelevant* to J1. Moreover, natural selection plausibly explains the smooth correlation, as stated in J1, much better than it explains fully determinate instances. Once a detrimental stimulus has been connected with displeasure sufficient to cause avoidance, there would not likely be further selection pressure to home in on an exact value. One might well expect that, in general, severely detrimental stimuli should be more unpleasant than milder threats, but again, most-determinate values would not be likely predictable from natural selection considerations.

Perhaps, then, determinate SBQ statements are evidence for interactionism or physicalism. But this is not right either. Physicalism says that experiences are physical. By itself, that does not imply any

[3] I used 'painful' rather than 'unpleasant' in giving James's version, because that is the term James uses. But Corabi uses 'unpleasant', and the argument's structure seems better served by this term. After all, many things that would be detrimental for us to ingest have unpleasant tastes and smells; but these are not properly called 'pains'. Hereafter, I will use 'unpleasant'.

SBQ statement, determinate or not. It just says that *whatever* SBQ relations we may find, the experiential component will be physical (and thus able to have effects). That is compatible with, e.g., arsenic being delicious. More generally: The falsity of a determinate SBQ statement is not evidence against physicalism. Parallel considerations hold for interactionism.

Perhaps determinate SBQ statements are evidence for the conjunction of interactionism or physicalism with natural selection. But, for reasons recently explained, natural selection plus either view will not predict determinate SBQ statements. What will be predicted, on either physicalism or interactionism plus natural selection, is that detrimental stimuli will generally be correlated with unpleasant experiences, and beneficial stimuli will generally be found pleasant. That is, they will predict J1.

J1 says nothing directly about behavior: it asserts the smooth correlation between utility and hedonic valence. So, perhaps we ought consider just the SQ part of SBQ statements. It will be evident on inspection that the preceding remarks apply just as well to determinate SQ statements.

Later in the same section, Corabi writes: "A good (albeit idealized) way to think of the confirmation process is to envision each general hypothesis (e.g., epiphenomenalism) as a disjunction of highly determinate versions of that hypothesis, each of which specifies the history of the world in maximal detail." But this does not seem to me to be a good way to think at all. The general hypotheses (physicalism, interactionism, epiphenomenalism) make no predictions about which determinate stimuli will be found correlated with which determinate qualia. Two of them predict that utility will smoothly correlate with hedonic valence, but none of them predict correlations between specific S and specific Q. Finding that there is a specific SQ correlation will thus not be helpful in deciding among the views.

Later in the paper, Corabi explicitly recognizes this point (in the section titled "The problem with the key assumption"). When we observe (or fail to observe) a particular determinate SQ correlation, none of the three hypotheses (physicalism, interactionism, epiphenomenalism) will gain or lose any ground. But if this is so — if determinate SQ correlations cannot be used to discriminate among rival views — then their claim to be the relevant form of "evidence"

is undercut.

Corabi's conclusion differs: it is that an improved version of the EA still fails. In my view, however, his revision of the EA is not an improvement. Instead, the insistence on maximum determinateness merely introduces irrelevant detail that obscures the heart of the EA. This heart is the smooth correlation stated in J1. Fragmenting this correlation into a plethora of determinate possibilities does nothing to strengthen the EA, and focusing on determinate characteristics rather than on the utility of stimuli and the hedonic valence of the experiences they cause removes the relevance of the "evidence".

3 Results for C1 – C3

The previous paragraph explains my rejection of C2 (the claim that the proposed "canonical" version is an improvement). Regarding C1, the case for "mishandling" of evidence rests on the fact that James is content with a general statement of smooth correlation, and does not require determinate SBQ statements. We have seen, however, that these statements are a distraction rather than an improvement, so we have no reason to accept that James mishandled evidence. On the contrary: J1 accurately states something that needs to be explained, and the sense of the argument is that efficacy for experiences better explains it than does epiphenomenalism. This remains a formidable argument, even for those who find grave difficulties in physicalism and interactionism, and even for those who may think that, when all the difficulties for the latter views are taken into account, the claim that the EA is a successful inference to the best explanation is undermined.

Regarding C3, it should by now be clear that James's version of the EA does not rely on any assumption about physicalism. So, C3 is false, as applied to it.[4]

[4] Neither does the EA say anything about physiological transitions in the brain (contrary to what Corabi says in the third paragraph of the section on "The central traditional confusion and the key assumption"). Any physiological hypothesis that is compatible with efficacy for pleasure and unpleasure will do.

4 Physicalism

Corabi's discussion of physicalism, in the section on "The problem with the key assumption", is couched in terms of "alternate determinate physicalist hypotheses". But at the end of section 2 above, we saw that these determinate hypotheses are not the right candidates for "evidence" regarding the EA. So, the problem that Corabi finds for the use of such hypotheses cannot show that there is a weakness in the EA.

However, I agree with Corabi that, even on the assumption of physicalism, statements of identity between properties that are specified in neural terms and properties that are not specified in neural terms are epistemically contingent (despite being metaphysically necessary, if they are true). And I agree that this fact presents a problem for those who wish to use the EA as an argument for physicalism. In what follows, I will briefly explain how epistemic contingency relates to the formulation of the EA given in the first section of these comments.

The focus of that formulation is pleasure and unpleasure. The claims of interest in discussing physicalism are thus claims of identity between physical (presumably neural) event properties, and pleasure or unpleasure. If we let 'NE1' and 'NE2' stand for the neural event properties that will (according to physicalism) be discovered to be identical with pleasure and unpleasure, respectively, we can express the crucial claims as follows.

P+ Pleasantness is identical with NE1.

P- Unpleasantness is identical with NE2.

Now, since these statements are not known to us *a priori*, it is intelligible to us that they might have been false (even though they are metaphysically necessary, if they are true). A plausible consequence of that intelligibility is that we cannot *explain* why the smooth correlation holds. For, as far as we can see, it might have turned out that NE1 was identical with pain, even while the laws of neurophysiology remained what they actually are. In that scenario, we would approach the painful, and doing so would enhance our fitness. Or, it might have turned out that alternative arrangements of brain parts other

than those involved in NE1 and NE2 would have established a condition in which the behavioral effects of NE1 and NE2 were reversed. In that case, if P+ and P- are true, our fitness would have been served by approach to the unpleasant and avoidance of the pleasant.

The Jamesian complaint is that epiphenomenalism cannot explain the smooth correlation. According to the line of thinking just described, physicalism cannot explain it either. So, if that line of thinking is correct, then physicalism is no better off in the face of the EA than is epiphenomenalism. Expressed in terms of the argument in the first section above, there is some reason to doubt J5.

The upshot of these reflections is this. We should agree with Corabi that the EA is not as powerful an argument for physicalism (or for [physicalism v interactionism]) as it is sometimes thought to be. But James's formulation of the argument remains formidable, because it does require a response from epiphenomenalists. They must either offer some alternative explanation of the smooth correlation, or cast doubt on the ability of rival views to explain it. James's formulation makes this burden clear. Reframing the matter in terms of determinate SBQ (or SQ) statements introduces irrelevancies that obscure such strength as the EA has.[5]

William S. Robinson
Iowa State University
402 Catt Hall
Ames, Iowa 50011-1306 USA
wsrob@iastate.edu

References

Corabi, Joseph. 2008. Pleasure's Role in Evolution: A Response to Robinson. *Journal of Consciousness Studies* 15:78-86.

James, William. 1890. *The Principles of Psychology*. Cambridge, MA: Harvard University Press.

Lipton, Peter. 2004. *Inference to the Best Explanation*. Second edition. London and New York: Routledge.

Robinson, William S. 2007. Evolution and Epiphenomenalism. *Journal of Consciousness Studies* 14:27-42.

[5] There is no rejection here of Bayesianism, properly applied. For explanation of compatibility of Bayesianism and IBE, see Lipton 2004.

Book Reviews

BIBLID [0873-626X (2014) 39; pp. 239-250]

Work and Object, by Peter Lamarque. Oxford: Oxford University Press, 2010, 248 pages. Pbk, ISBN 978-0-19-965549-6

The nature of works (of art) and their properties is *the* issue that unifies the many different questions addressed in *Work and Object*. It will be useful to present the view that Lamarque defends in this book by considering how it would apply to a particular work (of art). So consider Gaudí's *Casa Batlló*. The following is a description of the nature of this work of art and its properties according to Lamarque's view.

First of all, Gaudí's *Casa Batlló* is a *work* whereas the collection of glass, wood, iron, ceramics, Montjuic sandstone and the rest of materials that constitute it is a mere real thing or *object*. These two things, work and object, are distinct. They have different essential properties. For instance, while *Casa Batlló* is essentially a building, its constituting material is not. It could have constituted something different instead. And while Gaudí's work is imaginative, fantastical and original, its constituting material is not. Secondly, Gaudí's *Casa Batlló* would not have come into existence if the practice of architecture did not exist in the time of its creation and it would cease to exist if, and when, no one could understand architecture. *Casa Batlló*, like all the other works of art, was created, and hence brought into existence, when Gaudí completed it and decided it was complete under a conception of what had been achieved (i.e. a work of architecture). Thirdly, *Casa Batlló*, that is, the work, rather than its constituting object, is an intentional object whose nature cannot but be what is thought to be by our cultural human community, given that the very existence, nature and survival of *Casa Batlló* essentially depend on the practices and conceptions of that community. For instance, it is not possible for us to discover one day that *Casa Batlló* is not a building. *Casa Batlló*'s existence, though, does not depend on any individual mind. *Casa Batlló* is a real building, not an ideal object. It could exist and survive even if no one were to contem-

plate it, as long as the possibility of an appropriate appreciation of it remained. Fourthly, there is indeed an appropriate appreciation of *Casa Batlló*. For at least some of its properties are objective, even if response-dependent. These are at least its expressive and representational properties, such as *being moving* and *representing sea-life*. These properties depend on responses, but of well-informed specialized perceivers. Its purely evaluative properties such as *being beautiful* may not be objective, though. For they may depend on further specific conditions of its perceivers, well-informed and specialized or not. Fifthly, *Casa Batlló*'s vividness, for instance, is one of its aesthetic properties. It is an essential property it has, even if it is relational. On the one hand, vividness is relational in the sense that, as explained above, it depends on the response of a class of connoisseurs. And, hence, it involves a relation between *Casa Batlló* and a response of ideal perceivers. On the other hand, it is essential because nothing could be *Casa Batlló* unless it was vivid. Other aesthetic properties, the purely evaluative in particular, such as *Casa Batlló*'s beauty, may not be essential to it, though. Sixthly, *Casa Batlló* is dynamic, but it would not be dynamic, and would not have any other aesthetic property, unless its having those properties made a difference in a correct experience of it. Even an object that were as indiscernible from *Casa Batlló*'s constituting object as Danto's indiscernibles are from each other could be non-dynamic, or fail to share any other of its aesthetic properties, and thus, afford different (perhaps even perceptual) experiences. They would *look* different to well-informed appreciators. Knowing that *Casa Batlló* is one of Gaudí's creations, as well as knowing that it belongs to the category of architectural works, as Walton pointed out, is required for its correct appreciation and experience of its aesthetic properties. Seventhly, whereas *Casa Batlló* is a Catalan modernist building, and this is a general style recognized in features such as wavy lines and curved shapes that suggest natural forms, it also exhibits Gaudí's particular style, which is constituted by individual ways of implementing modernism. Gaudí's style is determined by Gaudí's psychological states, but this is not the case for the Catalan modernist general style, which is not based on any individual psychological state or process, but characterized by general features instead. Finally, *Casa Batlló* is emblematic and this property, among others, has been imputed to it through interpreta-

tion. After all, this is a property that *Casa Batlló* did not have since the very beginning, but acquired once it got interpreted.

As we could appreciate from this example, Lamarque argues that works (of art) are cultural intentional objects. In general, he characterizes works (of art) as having the following features: (a) they are real but distinct from the mere real things or objects that constitute them (*realism*); (b) their identity and survival conditions essentially depend on artistic practices and conceptions (or at least on appropriate reception conditions), though not on any individual mind (*non-idealism*); (c) they are created; (d) they are intentional objects in the sense that their nature *is* what is *thought* to be; (e) some of their properties are imputed by interpretation (*imputationalism*); (f) at least some of their properties are objective, depending on the responses of a class of ideal perceivers *(normativity);* (g) at least some of their aesthetic properties (if they have any) are essential, even if relational (*aesthetic essentialism*); (h) they have general as well as individual styles: their general style is well-captured by some of its characteristic features (*feature-based definition of style*), whereas their individual style is determined by the artist's individual ways of creating them and her psychological states and processes (*act-based definition of style*); and (i) they have different aesthetic properties only insofar as they afford experiential differences (*aesthetic empiricism*). In sum, this is what, according to Lamarque, characterizes works, both works of art and works which would not qualify as art.

As much as I share many of Lamarque's intuitions and I am very much sympathetic to the central tenets of this view, there are certain points in the book that did not convince me. Although there are many interesting issues I would like to comment on, I shall focus and structure my critical comments around two pervasive issues: what strikes me as a lack of a clear distinction between metaphysics and epistemology, and a lack of a consistent application of the distinction between work and object.

Starting with the first of these two issues, there are several ideas in the book which I think arise from not clearly distinguishing between metaphysics and epistemology. I will mention here just some of them. For instance, at some point in the book (p. 7) Lamarque discusses the question of the possible arbitrariness of ontology and easily concludes that ontology is arbitrary just on the basis of the

difficulties typically found in reaching a decision about what the best ontological theory is. This is to illegitimately get a metaphysical conclusion from epistemological premises. Clearly, the fact that it is hard to *know* what there is does not entail that there is no fact of the matter about what there is. Of course, this is compatible with some inescapable arbitrariness in the process of theory building. The theorist needs to make certain decisions that are arbitrary to some extent at some points in that process. But this arbitrariness is ultimately neutralized, when the theory faces its final test against the relevant data. Another example of unclarity about the distinction between metaphysics and epistemology is Lamarque's argument to the effect that aesthetics, unlike science and philosophy, cannot clash with common sense beliefs about works (of art). On his view, common sense beliefs about (art)works cannot be wrong precisely because a work (of art)'s origin, identity and survival *metaphysically* depend on human practices and conceptions. However, metaphysical dependence does not warrant epistemic infallibility. Not even if the dependence in question is on something related to epistemology, such as humans' conceptions. That works (of art) metaphysically depend on human practices and conceptions does not entail that humans cannot get them wrong. In this case, what we have is an illegitimate inference from a metaphysical premise to an epistemic conclusion. A third interesting case is provided by Lamarque's use of the Prehistoric cave paintings (p. 70, 115) as a clear example of his thesis that not only a work's creation and identity, but also its survival, depend on human practices and conceptions. Prehistoric cave paintings, he argues, are no work (of art), even if perhaps they once were, because we lack the appropriate knowledge required for a correct appreciation of them and this knowledge, it seems, can no longer be recovered. However, without further reasons I also fail to see this as a legitimate inference. My own intuition is that Lamarque's view is right with respect to the creation and identity of works (of art), that is, I think these do depend on the existence of human practices and conceptions, but not with respect to their survival. As the realist that I am, I can perfectly conceive of the survival of a work (of art) that is epistemically lost, and I have not found any other argument in the book for thinking otherwise. Lamarque assimilates survival conditions of works (of art) to those of things like legal facts and screwdrivers. But whereas

it is clear that the legality of same-sex marriage in Spain, for instance, would not survive the disappearance of the Spanish legal system, it is not that screwdrivers, or other artifacts, would not survive the disappearance of those who may use them, or that works (of art) would not survive the disappearance of those who may appreciate them. Further argument is needed. Finally, I would like to mention what I think is an ambivalent use of the word 'identity' throughout the book. This word seems to be used epistemically, while discussing metaphysical issues in which it should be used and understood metaphysically instead. An example of this, I think, is at those points in which Lamarque considers, somehow or other, Kripke's metaphysical thesis of the essentiality of origin in its application to works (of art). Like when he argues (p. 72) that the origin of certain works is not essential to their *identity*. According to Lamarque, origin is not essential to (the identity of) all works because knowledge of their origin is not required for a correct appreciation of all of them. Jingles, minor pop songs and football chants are among the examples he provides of works whose origin is not essential. But one thing is that knowledge of the origin of something is irrelevant to its appreciation, and yet another that its origin does not determine its identity *metaphysically*. Something similar happens in the book's discussion of the possibility of the so-called 'referential forgery of allographic arts' (p. 83). Metaphysical origin-related conclusions about the *identity* of physical tokens of literary and musical works (of art) seem to be derived from questions about their value and importance. It is precisely because the provenance of type-instantiating text-tokens has little value that a scenario like the one that Lamarque discusses of the multiple production of type-instantiating text-tokens of White's poem and Black's (identically worded) poem is possible. In this scenario intentions, or external factors, determining provenance are rather weak and this is what makes it less determinate. That a text-token may be read as different works (of art) and that how to read it may be decided by the reader (because nothing really hinges on this) does not mean that provenance does not determine what work (of art) the text-token is a copy of. The fact that provenance of particular tokens is important for the appreciation of paintings and sculptures and not for the appreciation of literary works has to do with the fact that paintings and sculptures (or their constituting objects) are physical

particulars too, whereas literary works (and the objects that constitute them) are types rather than physical text-tokens. The physical text-tokens that we usually use are mere means to access the work (of art), and even the object that constitutes it. If this is right, none of this shows that provenance does not determine (the identity of) a work (or a token) in the metaphysical sense. And neither it shows that referential forgery of non-particular musical and literary works (of art) is not possible. Even if provenance is less important in the case of type-instantiating tokens, and the scenario considered is one in which provenance of the relevant tokens is complex, referential forgery may still be possible by producing a type-instantiating token of a work X while presenting it as a type-instantiating token of a different work Y, in the way that Levinson argues. Again, of course, referential forgery of allographic arts does not matter much, and this is what makes it special in the sense of less interesting. Here again Lamarque does not seem to distinguish between the nature of something and (the importance it has on) how it is taken to be.

I shall now move on to the second issue I announced above, that is, the lack of a serious and consistent application of the distinction between work and object. One instance of that is in Lamarque's recurrent use of the obscure and controversial *qua*-talk. On the one hand, taking the distinction between work and object seriously makes this *qua*-talk unnecessary. This is because for every work (of art) there are two things with different properties: i.e. the work itself and the object that constitutes it. So it is superfluous to talk of works or objects *qua* works or *qua* objects, or of them *under certain descriptions*. On the other hand, I think that all of this *qua*-talk has already been justifiably discredited, at least in certain philosophical contexts. Things have properties *simpliciter*, not *qua* anything else or *under certain descriptions*. Take the Superman story (discussed by Lamarque in p. 62), and consider it as if it were factual rather than fictional. Many of us agree in that, for instance, Lois Lane believes that Superman is sexy whereas she does not believe that Clark Kent is sexy. However, as any of the main theories of propositional attitudes and their ascriptions would show, this does not mean that Superman has a property (i.e. *being believed by Lois Lane to be sexy*) that Clark Kent does not have. This would not make sense. There is only one individual, Superman *is* Clark Kent. And so, it is this

individual, call him 'Superman' or 'Clark Kent', who has that property. This is wholly compatible with the truth of the belief reports mentioned above, as long-standing work on propositional attitudes attributions has clarified. Likewise, it does not make sense to say that Superman *qua* the superhero (or *under the description of the superhero*) can fly whereas Superman *qua* the reporter (or *under the description of the reporter*) cannot fly (even if Superman only flies when dressed-up as a superhero). The case of Jones, the mayor, is analogous (p. 48). As a mayor, Jones has some office-related duties that he also has as a man while being a mayor. One thing is to think of Jones *qua* mayor in the sense of considering Jones as a mayor and yet another to metaphysically distinguish Jones-the-mayor (or Jones *qua* mayor) and Jones-the-man (or Jones *qua* man). There are no two such things with different properties, and no single thing with different properties either. In any case, as I said above, all of this talk is unnecessary because, unlike these two cases, the case of works (of art) do involve two different things with different properties. We only need to take the distinction between work and object seriously. The distinction between person and role (p. 106) does not parallel the distinction between work and object either. Against what Lamarque claims, it is not true *that the British monarch is head of the state* is a necessary truth. This is not an essentialist claim that rests on an implicit *qua* operator qualifying the subject term, in the sense that it is only *qua* monarch that the British monarch is head of the state. Instead, what would be a necessary truth, if certain British political facts could not change, is *that whoever, if any, is ever the British monarch is head of the state*. The role of being the British monarch is not itself any head of the state, but any person who ever occupies this role contingently may be. A final example of this kind of controversial talk that I shall mention occurs in discussing style. Lamarque talks of stylistic properties under an act-based definition as properties that acts have only *under a description*. As an example he mentions that a certain kind of movement would be graceful (in style) if it were a *dancing* movement, whereas it would be ungainly (in style) if it were a *running* movement (p. 141). He seems to be thinking about properties that acts have in virtue of their being the things they are (or in virtue of belonging to the categories they belong). A dancing movement that is graceful is a dancing movement, not a running movement, and it is graceful regardless of

the description under which it is considered, even if its gracefulness depends on its being the thing it is: that is, a dancing rather than a running movement. The same applies to running movements and their styles. Well understood, this seems to amount to nothing more than Walton's well-known point as applied to the case of style: that is, that the aesthetic properties that something has depend on the category it belongs to.

There are many other interesting issues that Lamarque addresses in *Work and Object* that I would love to but cannot discuss within the limits of this review, such as his defense of imputationalism and his conception of fictional characters as interest-relative types. But I would not like to finish without dedicating a few lines to Lamarque's appealing thoughts on conceptual art in the last chapter of the book. According to Lamarque, urinals, Brillo boxes and the like could only constitute works (of art), even if works of *conceptual* art, as long as the artist manages to create something different from them. And this happens just in case there's something other than the ordinary object that according to Lamarque invites a kind of perception, which makes salient particular aspects and suggests significance for them. That extra thing materially constituted by these ordinary things is the work. That work is not just ideas, a physical medium acting as a vehicle for the transmission of ideas is an important part of works of conceptual art too. The ideas must inform the perception of these works. Perceiving that vehicle, or a copy of it in the case of works that are types rather than particulars, seems crucial for a correct appreciation of the work. For there are correct and incorrect ways of responding to a work of this kind too. In conceptual art, subjective responses are correct, while the search for any single or true interpretation is not. After all, works of conceptual art are intended to generate reflections on ideas.[1]

Gemma Celestino
LOGOS – Research Group in Analytic Philosophy
University of Barcelona
C. Montalegre, 6
08001 Barcelona

[1] Research leading to this work was partially funded by the research projects FFI2011-29560-C02-01 and CSD2009-00056.

Spain
gcelestino@ub.edu

Art and Art-Attempts, by Christy Mag Uidhir. Oxford, Oxford University Press, 2013, 222 pages.

What do Ovid, Dante, Petrarch, Camões, Cervantes, Shakespeare, Caravaggio, Velázquez, Rembrandt, Bach, Goya, Mozart, Beethoven, Turner, Hugo, Tolstoy, Eliot, Pessoa, among others, have in common? One answer is simple: they all have been the creators of great works of art. But what makes something a work of art? What is art? Here the puzzles begin and the philosophy of art attempts to answer these and related questions. The meta-philosophy of art seeks to provide a framework in which these questions can be addressed.

In *Art and Art-Attempts*, Christy Mag Uidhir aims at providing such framework. He begins with the assumption that art is "intention-dependent" and he investigates "what follows from taking intention-dependence seriously as a substantive necessary condition for being art" (p. 6). This he calls the 'Attempt Theory of Art'. As he warns the reader, the Attempt Theory of Art "is *not* itself a theory of art" (p. 6), but what we might call a meta-theory: it focuses on what a theory of art must be, minimally, to be viable as such. The purpose is not to enquire into the nature of art, but to provide "something even better: a unified, systematic, and productive framework for philosophical enquiry into art" (p. 209).

The first chapter is crucial and it deals with "art and failed art". Mag Uidhir never spells out the conditions for something being art (he begins by professing ignorance about this) but he claims that "the way in which [a] thing comes to satisfy the conditions for being art (whatever those may be) must be the product of intentional action" (p. 23). (He purports to begin with an assumption that is uncontroversial.) Here he gives an example that shows that his Attempt Theory, rather than being unanimously accepted as he claims, is quite controversial. He asks us to imagine that he attempts to paint a realist portrait of his aunt Teresa. Since he is an "inept painter" and the result does not resemble his aunt "in the slightest", he fails to produce a portrait of his aunt Teresa. With this everyone agrees.

"However, the irregularly shaped blob possesses rather striking aesthetic properties, though only as an accidental (and unbeknownst to me) result of actions intended to be in service to the portraiture" (p. 34). Mag Uidhir concludes that the result is *not* a work of art (it is *failed art*) because the aesthetic properties that the work does possess are not the result of the intention to produce them: "the work has those properties as the result of the way in which my attempt at portraiture *failed* and not as the result of any successful art-attempt" (p. 34). So, he concludes, even though the work does possess aesthetic properties, and *appears* to be an artwork, it is not one. He says it is "complex failed-art" (p. 35) and he adds that "it could be the case that many things thought to be art are in fact complex failed-art" (p. 35). Indeed, if all aspects with aesthetic interest need to be intended in order to be artistic, then there are many works that are in fact complex failed-art according to Mag Uidhir.

This example is illustrative of the controversial aspect of the Attempt Theory, despite Mag Uidhir's claims that the theory is acceptable to all. While professing ignorance about the nature of art (p. 1), Mag Uidhir claims that it is not sufficient for a work to be art that it has aesthetic properties: they need also be the result of intentional actions of the appropriate type. So despite his attempt to remain neutral with regards to theories of (the nature of) art, Mag Uidhir's tacit theory rejects at least the aesthetic theory of art, a theory which could give art status to his failed portrait of aunt Teresa. Moreover, it does not seem true that his Attempt Theory applies to all works of art, even though it applies to many. For instance, Anne Frank's *Diary* and Fernão Lopes's *Chronicles* were not literature attempts (and therefore not art-attempts), but both are now regarded as literature and therefore as art. Anne Frank's *Diary* was meant to be a journal, with no literary pretensions, and Fernão Lopes's *Chronicles* were attempts at history, even though they are now studied in Portuguese Literature courses and read as literature: their aesthetic (more precisely, literary) properties have made them gain that art status.

So Mag Uidhir's "Attempt Theory" and his distinction between art and failed art is far from being uncontroversial, and despite the author's claims of independence from any substantive theory of art, it relies on a tacit theory of art that is at least a rejection of an aesthetic theory of art. In fact, the Attempt Theory which Mag Uidhir

puts forward denies artistic status to the aesthetic properties that were not intended by the artists. This seems to me a high price to pay, since it is certain that artists are not aware of or responsible for all the aesthetic properties their works end up possessing: many aesthetic effects are unconsciously produced and part of the value of works of art is due to their amazing and surprising aesthetic results, arrived at in a variety of ways, consciously to a great extent but with unconscious elements as well. Furthermore, it is not always clear (in fact, most of the time it is not clear) which aesthetic features were intended when producing a work of art, so Mag Uidhir's theory leaves us with uncertainty as to what we can interpret as artistic in most works of art. In addition, the intentions of the artist are not always publicly available. So if Mag Uidhir's theory is correct we can say very little (from an artistic point of view) about most works of art.

The theory that "something is an artwork if and only if that thing is the product of a successful art-attempt" (p. 86) is, therefore, far from uncontroversial. Moreover, without endorsing a theory of art it is hard to see what distinguishes a failed art-attempt from a successful art-attempt. Mag Uidhir's theory requires additional clarification about what makes art count as art. Saying "whatever those [conditions] may be" (*passim*!) is not enough. For instance, we need to know what makes Camões's epic poem a masterpiece. We need a principled way of distinguishing between Júlio Dinis's largely failed attempts at poetry and his clearly successful attempts at novel writing. We need to know what makes Eça de Queirós's novels so successful as literature. (We certainly don't want to claim that public, institutional, success is the only criterion to distinguish good art from bad art or failed art).

The controversy of the Attempt Theory does not end here. For example, the wish to preserve the Attempt Theory leads Mag Uidhir to conclude that "PHOTOGRAPHY cannot be an art form because to be a photograph is to be the mere causal product of a certain photochemical process, and being a mere causal product of photochemical processes is neither attempt-dependent nor intention-dependent" (p. 119). Granted, not all attempts in photography are works of art. But neither are all attempts in painting or literature or music always successful artistic attempts.

Despite my disagreement, I must say that this book is an inter-

esting and praiseworthy book. With a variety of examples, most of them from contemporary art, *Art and Art-Attempts* offers thought-provoking discussion in the meta-philosophy of art. The choices of the artistic examples assume, however, a theory of art that is more inclusive than some readers would be prepared to endorse. And Mag Uidhir does not tell us what are the criteria used to endorse his tacit theory of art. So he leaves us with no way of distinguishing art from "failed art". For example, Marcel Duchamp's *Fountain*, John Cage's *4'33''*, Tracey Emin's *My Bed* and (*pour couronner le tout*) Manzoni's *The Artist's Shit* are all art-attempts. But what makes them "successful" or "failed"? Is the admission into the circle of art critics sufficient for a work to qualify as a successful art-attempt? The problem is that Mag Uidir's account does not provide a way to decide on this crucial matter, leaving us with no distinction between works of art and failed works of art. He offers the following:

> artworks and failed-artworks are both products of the right sort of attempts, the difference being that artworks satisfy the conditions for being art (*whatever those may be*) by virtue of the way in which those attempts succeeded while failed-artworks do not satisfy the conditions for being art (and so, are non-art) by virtue of the way those attempts failed. (p. 17)

The underlying and professed ignorance on this matter is therefore crucial. To give the art examples Mag Uidhir gives, he must endorse a theory that allows their inclusion.

We are thus left very curious about what makes some things works of art and others just "failed" works of art. This book is very thought-provoking and a good contribution to discussion in the philosophy of art. The starting point and main thesis is, however, and despite the author's claims and best intentions, contentious and will certainly invite questions and rebuttals.

Inês Morais
ibmorais@gmail.com