Croatian Journal of Philosophy

Vol. XVIII · No. 54 · 2018

Croatian Journal of Philosophy

1333-1108 (Print)
1847-6139 (Online)

Co-published by
"Kruzak d.o.o."
Naserov trg 6, 10020 Zagreb, Croatia
fax: + 385 1 65 90 416, e-mail: kruzak@kruzak.hr
www.kruzak.hr

and

Institute of Philosophy
Ulica grada Vukovara 54/IV, 10000 Zagreb, Croatia
fax: + 385 1 61 50 338, e-mail: filozof@ifzg.hr
www.ifzg.hr

Available online at http://www.ceeol.com and www.pdcnet.org

CROATIAN JOURNAL OF PHILOSOPHY

Vol. XVIII · No. 54 · 2018

Articles

Logic of Argumentation

Book Discussion

Book Reviews

Croatian Journal of Philosophy
Vol. XVIII, No. 54, 2018

An Anscombean Reference for 'I'?[1]

ANDREW BOTTERELL and ROBERT J. STAINTON
The University of Western Ontario, London, Canada

A standard reading of Anscombe's "The First Person" takes her to argue, via reductio, that 'I' must be radically non-referring. Allegedly, she analogizes 'I' to the expletive 'it' in 'It is raining'. Hence nothing need be said about Anscombe's understanding of "the referential functioning of 'I'", there being no such thing. We think that this radical reading is incorrect. Given this, a pressing question arises: How does 'I' refer for Anscombe, and what sort of thing do users of 'I' refer to? We present a tentative answer which is both consistent with much of what Anscombe says, and is also empirically/philosophically defensible.

Keywords: G. E. M. Anscombe, 'I', persons, immunity to error through misidentification, deflated reference, The First Person.

1. *Introduction*

Our goal in this paper is to extract a novel reading from G. E. M. Anscombe's classic paper "The First Person" and to defend the view that we take her to hold. This is no easy feat, since much has been written about that paper—and much of that has been negative. But we believe that there is an overlooked reading of "The First Person" that is both

[1] A first draft of this paper was presented at the Department of Philosophy, Universidad Alberto Hurtado, Santiago, Chile, May 24, 2017. A revised version was given at the Annual Conference on Philosophy and Linguistics, Inter-University Center, Dubrovnik, Croatia, September 5, 2017; to the School of Philosophy, Australian National University, November 2, 2017; to the Department of Philosophy, University of Reading, January 19, 2018; and to the Department of Philosophy, University of Manchester, May 22, 2018. We are grateful to audience members at these venues for very helpful feedback. Thanks also to Brian Garrett, Jennifer Hornsby, Léa Salje and Barry Smith for very pleasant mealtime discussions of Anscombe's views, and to Lenny Clapp, Rick Grush, Michael Hymers, Rockney Jacobson, Angela Mendelovici, Martin Montminy, and Eliot Michaelson for written comments. Financial support was provided by The Social Sciences and Humanities Research Council of Canada and by a Visiting Fellowship from the Research School of Science Sciences at the Australian National University.

consistent with much of what Anscombe says there and elsewhere and empirically/philosophically defensible.

First some stage-setting. There is a fairly standard reading in the literature on "The First Person" according to which Anscombe is arguing that the first-person pronoun 'I' is radically non-referring. On this "Straight" reading, far from functioning logically as a proper name, 'I' is instead, for Anscombe, similar to the syntactically expletive use of 'it' in 'It is raining'. So read, her core argument is that 'I' must be non-referring since otherwise we arrive at a metaphysical view such that something like a Cartesian ego exists and is the referent of tokens of 'I'. And according to Anscombe, that view is borderline nonsensical. Philosophers who read "The First Person" in this way include Clarke (1978), Evans (1982), Garrett (1994, 1997), Hamilton (1991), Hinton (2008), Kripke (2011), Peacock (2008), Taschek (1985), Teichmann (2008), Wiseman (2018) and van Inwagen (2001).[2]

In a recent paper, one of us has defended a revisionist alternative to this Straight reading of "The First Person" (Stainton 2018). Goes the idea, Anscombe can be seen to be making at least three points. The first is that 'I' doesn't behave like a proper name *as proper names were understood at the time.* The second point, as we in 2018 might phrase it, is that in one *historically specific sense of 'refer'*, 'I' doesn't "refer". Rather, 'I' can be used to "speak of" something (47); to "concern an object" (61 and 63); and to "specify" an object (47).[3] The third is that when thinking about 'I', we should not be misled by surface grammar. For while 'It is raining' has a surface-subject term, in that context 'it' is non-referring (on every construal of 'refer'), and contributes nothing to the sentence's meaning.

Our question begins where this revisionist account leaves off. In short: if the Straight reading of "The First Person" is rejected, *how* on Anscombe's view does 'I' manage to "speak of", "specify" or "concern" things in the world? And *what* does one "speak of" when using 'I'? To put it deliberately vaguely for now, so as not to beg any questions, our focus will be: What, to use her phrase, is the "mode of meaning" of 'I' for Anscombe (55)?

[2] One example: "Professor Anscombe's position is that it is not the function of the word 'I' to refer; the word is thus unlike 'the present kind of France', which is in the denoting business but is a failure at it; rather, the word, despite the fact that it can be the subject of a verb or (usually in its objective-case guise, 'me') the object of a verb, is not in the denoting business at all… for Anscombe, the word 'I' refers to nothing in a way more like the way in which 'if' and 'however' refer to nothing" (van Inwagen 2001: 6). Reading Anscombe in this (standard, widespread) fashion, van Inwagen takes her view to be easily refuted: e.g., by the logical validity, due to transitivity of identity, of 'I am Elizabeth Anscombe; Elizabeth Anscombe is the author of *Intention*; therefore, I am the author of *Intention*'. To our minds, the utter obviousness of such an objection shows that this cannot really have been Anscombe's position on 'I'.

[3] Unless otherwise noted, all in-text citations are to Anscombe (1975).

One might be excused for wondering: Why expend so much time reconstructing, from Anscombe's text and from her larger body of philosophical work, a positive story about 'I'? In other words, why Anscombe? And why "The First Person"? Our motivation involves a mix of the historical and the substantive. The historical motivation is that such a reconstruction encourages renewed engagement with her extremely original and important work on language and mind. Anscombe is surely one of the greatest philosophical minds of the 20th Century, whose work on action theory and ethics is foundational for entire subfields. As a result, her oeuvre surely merits the same scholarly respect as that devoted to many of her male peers in the Analytic tradition, such as Austin, Davidson, Dummett, Grice, and Strawson. And one way to illustrate the importance of Anscombe's work in mind and language is by engaging directly with this underappreciated paper of hers, one that has spawned a huge literature and is jam-packed with insights—albeit ones often denigrated as fruitful ideas that appear in the context of a not-very-convincing paper.

Those are our historical motivations. Substantively, we find in "The First Person" an initially promising view about the semantics of the first-person pronoun 'I'—one worthy of further development quite independently of Anscombe's historical standing in the field. In short, even if Anscombe were not one of the founders of the Analytic tradition, her insights and arguments in "The First Person" would still be worth taking seriously.

So much by way of stage-setting; here is our plan going forward. We begin with methodological remarks. Next, we explicate some Anscombean observations about 'I' that any successful account of its "mode of meaning" must accommodate. We then present our positive view: in particular, we will attribute to Anscombe the insight that 'I' has (what we call) a "deflated reference". We then argue that this view is plausible both as a tentative piece of Anscombe exegesis as well as a substantive proposal about the syntax and semantics of 'I'. We conclude with some objections and replies.

2. *Methodological Preliminaries*

Our twin motivations lead us to adopt a certain methodological approach: a sort of history of philosophy that lies between two poles. It is not philosophy-focused history nor is it historically-inspired philosophical problem solving. Our neither-fish-nor-fowl methodology yields twin criteria for success. First, the better our reconstruction fits with the text and with the author's larger corpus and philosophical milieu, the better the reconstruction. Second, the more promising the reconstruction is *qua* substantive account of the phenomenon, the better it is.

These two criteria are potentially conflicting. One would like to be charitable to the author, but one doesn't want to be too charitable. Great philosophers get things wrong and we certainly acknowledge that Anscombe's work, both in "The First Person" and elsewhere, is imperfect: her writing style is often obscure and she is sometimes too dismissive of opposing or conflicting views. Relatedly, a "perfectly charitable" reading threatens to be anachronistic. So, our approach requires balancing out "what Anscombe really thought in the early 1970s" against "what we can learn from her about our present-day issues."

Putting a positive spin on this difficult balancing act, both poles stand to benefit from a satisfying answer to our target question; it could provide a useful departure point for each. It would also provide indirect support for the conclusion of the companion negative paper mentioned above: that Anscombe eschews the radical non-referring view becomes all the more plausible if our hypotheses herein are on the right track.

3. *Anscombe on 'I'*

Moving beyond methodological commitments, we turn to some core elements of "The First Person". Anscombe's free-flowing style resists regimentation, but many authors would agree that her positive remarks about the way 'I' functions can be distilled into a handful of observations.

Observation #1: Immunity to Certain Errors

According to Anscombe, 'I' seems to be immune to reference failure: "If 'I' is a name, it cannot be an empty name" (55). 'I' appears equally immune to a certain kind of error regarding mistaken identification: "Guaranteed reference [in this latter sense] would entail a guarantee, not just that there is such a thing as *X*, but also that what I take to be *X is X* (57). Or again: "[The 'I'-user cannot] take the wrong object to be the object he means by 'I'" (57). Here is an example designed to support those generalizations:

Rob: I am smoking
Andrew: #You're right that someone is smoking, but the person you intended by 'I' is actually Juanita, not Rob

'I' in the first sentence cannot fail to refer. (Or so it seems. The point will be revisited below.) This contrasts with, for example, the expression 'The man with the hat' in 'The man with the hat is smoking'. A speaker, say Rob Stainton, could use it when looking at what is in fact a trick of the light, and thereby fail to refer to anything. More intriguingly, the absurdity of the second sentence highlights that a speaker cannot wish to refer to one thing with 'I' and yet somehow end up referring to something else. Again, contrast 'The man with the hat is smoking'. It is perfectly possible for Rob Stainton to use it to pick out, and talk about, a woman with a large, geometrical hairdo; and a perfectly sensible reply could be 'You're right that someone is smoking, but the person you intended by 'The man with the hat' is actually a woman with a curious head of hair'.

To come at the point another way, 'I'/'myself' seems to have an "indirect reflexive" use such that *I spoke of myself, but I didn't know it* is not possible. On that use, someone saying 'I' cannot misidentify the referent—so, such a confusion cannot arise. Here again, this is to be sharply contrasted with the "direct reflexive" in 'When I spoke of the man with the hat, I spoke of myself, but didn't know it', where such misidentification is perfectly possible.

Observation #2: Immunity to Doubt

The foregoing facts about immunity-to-referential-error also yield epistemological consequences. Though I (that is, Rob Stainton) can doubt whether Rob Stainton exists, thinks, and so on, I cannot doubt whether *I* exist, think, and so on. Relatedly, while I can doubt whether I (that is, Rob Stainton) am Rob Stainton, I cannot doubt (in the "indirect reflexive" use) whether I am me. So 'I'-talk seems to rule out certain skeptical worries.

Observation #3: Bodily Properties

A third Anscombean observation is that if 'I' refers, then one can conceive of it doing so in the absence of a body altogether, or indeed in the absence of any bodily sensations. In support of this view, Anscombe introduces a much-discussed Tank Thought Experiment: in an imagined situation of utter sensory deprivation, urges Anscombe, a person can still think: 'I won't let this happen again' (58). To support the same conclusion, she proposes a Body-As-Puppet Thought Experiment. The following sentence, suggests Anscombe, could be used and understood in a conceivable conversation: 'When I say 'I', that does not mean this human being who is making the noise. I am someone else who has borrowed this human being to speak through' (60). Here, it does seem that what 'I' would refer to need not be any kind of physical body.

Observation #4: Perception and Action

Finally, the observation that, say, the man in the hat is in danger (where the man in the hat is, as a matter of fact, Andrew Botterell) can have quite different action-generating effects than the observation that *I* (that is, Andrew Botterell) am in danger. Closely related to this, 'I' can be used to express an intention to act in a certain way. This is very different from using 'I' to make empirically-based predictions about how a certain body (for example, that of Andrew Botterell) will behave in the future (56).

By way of summary, contrast the name 'René Descartes'. It lacks many of the foregoing features. For example, the following discourse makes perfect sense:

> Rob: René Descartes is smoking.
> Andrew: You're right that someone is smoking, but the person you intended by 'René Descartes' is actually Baruch Spinoza, not René Descartes.

This shows that proper names are not immune to certain sorts of reference-errors. The same can be said about immunity to doubt: while René can doubt whether he is René Descartes, he cannot doubt that he is himself (in the "indirect reflexive" sense).

Moreover, as we read her, Anscombe would *disagree* that 'René Descartes' might refer to a disembodied soul or Cartesian ego. As we will discuss below, Anscombe thinks that it's built into the meaning of 'Chicago' that it refers to a city; similarly, it's built into the meaning of 'René Descartes' that it refers to some sort of embodied animal, specifically a human male.[4] Finally, the action-generating effects of 'René is in danger' are comparable to those of 'The man in the hat is in danger'; they aren't like those of 'I am in danger'.

In addition to these four positive observations, "The First Person" contains several negative points about how *not* to account for them. First, Anscombe notes that attributing to 'I' a special "descriptive sense" (in the Frege-inspired sense) won't do the trick. This holds even if what is proposed is a sense that is merely envisioned by the speaker: e.g., a sortal intended by the speaker to fix the referent of the bare demonstrative 'this'. In particular, according to Anscombe, one must not assign a descriptive sense to 'I' that would lead to a Descartes-type mentalistic "self" being the referent of 'I', such that: I have infallible knowledge of that mental "self"; aspects of it are "private" in that only I can have knowledge of those; and the "self" is made of some queer non-bodily substance that explains these properties. Anscombe also warns that one should not attempt to ensure guaranteed reference by having the pronoun pick out only the me-right-this-instant. Rather, 'I' must be capable of specifying entities that have a temporal extension.

It is on the basis of these arguments and observations that Anscombe (in)famously concludes with the seemingly extraordinary claims that have animated the Straight reading:

(i) Logically speaking, 'I' is not a name (53 and 56);
(ii) I' does not involve singular reference (53);
(iii) 'I' does not refer to the 'I'-user (56):
(iv) 'I' is not a singular term whose role is to make a reference (56 and 58);

[4] To anticipate, this may prove one part of the reason why, in Anscombe's view, 'I am Elizabeth Anscombe' is *not* an identity proposition: we will urge that, for her, 'I' specifies a person in the forensic sense; and that person is (as one might variously put it) merely connected with/realized by/composed of a living human male body. If the person-*qua*-moral-agent and her body are not one and the same thing then (even though 'I' is used to "speak of" things, hence not an expletive), Anscombe's infamous claim about 'I am Elizabeth Anscombe' looks reasonable.

(v) 'I' is neither a name nor another kind of expression whose logical role is to make a reference, *at all* (60)

4. *On 'Referring'*

Recall our deliberately vague target question: What, according to Anscombe, is the "mode of meaning" of 'I'? How might we go about answering this question? We think it is best to proceed in stages. First, recognize that it is only "inflated" reference that is being rejected by Anscombe. Second, identify an alternative that fits better with her text and larger philosophy.

To begin with, it is clear that the then-current Frege-inspired conception of reference builds in a great deal. It requires that proper names have a descriptive sense that is synonymous with a definite description. Empty names aside, that sense fixes a substantial objective thing as the referent of a name.[5] It also licenses various *a priori* entailments and analytic necessities (e.g., the descriptive sense of 'Chicago' *a priori* entails that it is a city; and this is, as a matter of meaning, a necessary feature of Chicago). To elaborate with a notorious example, assuming 'Hesperus' has as its descriptive content *first heavenly body visible at night*, this descriptive content would simultaneously fix the referent as Venus and make the name synonymous with the noun phrase 'The first heavenly body visible at night'. As a result, it will be analytic that Hesperus is a heavenly body; anyone who knows the meaning of the name will know *a priori* that this is the case; and the heavenly-body status will be necessary. Reference (of this "inflated" variety) also requires, second, that the speaker *intend* a descriptive content (56): typically, this will coincide with the descriptive content of the term, although the speaker may unwittingly intend a different content, thereby fixing a different "speaker's referent". Turning now from the reference relation to the thing referred to, an "inflated referent" must, third, be a "distinctly identifiable"/"distinctively conceived subject" (65) having clear identity conditions (53). Finally, the required descriptive content and the required "objective/robust" nominatum jointly explain not just epistemological and metaphysical features of the term, but also psychological ones: e.g., that perfectly rational agents can fail to realize that Hesperus is Phosphorus is explained thereby.

[5] Textual evidence that Anscombe demands a conception/sense for "name-like words" and for "reference" (as she uses those terms), includes: "We seem to need a sense to be specified for this quasi-name 'I'. To repeat the Frege point: we haven't got this sense just by being told which object a man will be speaking of, whether he knows it or not, when he says 'I'… [If] 'I' expresses a way its object is reached by him, what Frege called an "Art des Gegebensein", we want to know what that way is and how it comes about that the only object reached in that way by anyone is identical with himself" (48), Also: "The use of a name for an object is connected with a conception of that object. And so we are driven to look for something that, for each 'I'-user, will be the conception related to the supposed name 'I'..." (51–52).

On the view that we are extracting from "The First Person", 'I' does not exhibit reference of this "inflated" sort. 'I' has instead only a "deflated reference", in at least three senses: a deflated referring relation, a deflated referent, and a deflated psycho-philosophical explanatory burden. Let us unpack these, each time taking philosophically inspiration from other authors.

A. *Deflated Referring Relation*

To explain what we have in mind when we talk about 'I' having a deflated *referring relation*, we borrow from David Kaplan's work on "pure indexicals". According to Kaplan (1989), such terms have no descriptive sense associated with them. Rather, they obey a rule-of-use that outputs an object given a context of utterance. Importantly for our purposes, pure indexicals (including 'I') don't invoke the intentions of the speaker. Anscombe herself phrases the rule-of-use for 'I' thus: "If *X* makes assertions with 'I' as subject, then those assertions will be true if and only if the predicates used thus assertively are true of *X*" (55).[6]

B. *Deflated Referent*

The rule-of-use proposed above requires that there be *something* that a token of 'I' concerns or specifies. Critically, the rule does not itself fix whether that something is a soul, a mental substance, a body, etc. It merely says that the thing-asserting, whatever it be, is what will make the assertion true or false. As we read Anscombe, it will be facts about our world that settle which things turn out to be assertion-makers hereabouts.

Anscombe clearly does not believe that assertion-makers are chunks of Cartesian inner mental substance. She eschews any such thing as nonsensical. But then what can be the deflated alternative? What else, for her, can stand in for *X*?

We can find something suitably Anscombean if we move away from a preoccupation with a Descartes-inspired mentalistic "self" and towards something very different. An important kindred spirit, we think, is Peter Strawson (1953, 1959, 1966). According to him, and putting things crudely, there is a gradient among "individuals" running from the most primitive proto-individuals with mere feature-placing (e.g., raining or smelling foul hereabouts) to the most robust—countable, clearly individuated, self-standing, and explanatory objects (e.g., the dog Fido). Crucially for the positive view that we are reconstructing, and consonant with Anscombe's philosophical *foci*, along this continuum there can be individuals that are a (mere?) *locus of ethical evaluation and*

[6] More cautiously, and as Anscombe herself explicitly recognized in her *Post Scriptum* at p. 65, because of the existence of "oblique" contexts this proposed rule-of-use for 'I' would need to be revised somewhat. It should read something like: '… those assertions will *ordinarily* be true if and only if…'. Oblique contexts would then be treated as non-ordinary exceptions. More on this below.

intentional action: persons in the "offenses against the person" sense, to use Anscombe's well-fitted phrase (61). So understood, persons are very unlike the philosopher's mind-internal "selves": persons are not distinctly identifiable subjects whose queer nature (causally) explains the emergence of normatively evaluable actions. Nonetheless, we are suggesting that they *are* (intersubjectively observable) "objects" that one can straightforwardly talk about—indeed, in the usual case, the sorts of things that exhibit the features of ethical evaluation and rational action are, for Anscombe, living human bodies (61).

A related insight can be found in Amie Thomasson's writings (see, e.g., her 2010). An important line of thought therein is that the ontological scruples of Quine (1948)—which require precise individuation conditions, reducibility to the physical sciences, etc., before something can be counted as a genuine object—are overly demanding. To the contrary, many perfectly respectable entities fail to meet such arch conditions: silences, holes, storms, academic disciplines, Nominalism, folk songs, and so on. These too would all be, in our sense, "deflated referents". These ideas apply to "The First Person" in the following way: a referent for 'I' need not have precise identity conditions. Instead, what it is for there to be an individual, "the person", for which the first-person personal pronoun 'I' can stand, is merely for there to be something-or-other that acts rationally, and that is subject to normative evaluation. Relatedly, to demand that the existence of persons, in this forensic sense, explains how there come to be normatively evaluable actions gets things the wrong way around. (Compare: "Rules are prior to and explain behavioral patterns and (in)correctness". No, says the Wittgensteinian, it is because there are behavioral patterns and (in)correctness that it's proper to recognize a rule.)

One should identify the referential locus of 'I' as the person, forensically understood, not merely because of persons' centrality to action theory and ethics, but also because, as Anscombe says, "only thoughts of actions, postures, movements and intended actions… are unmediated and non-observational" (63). Coming at things this way, one can take Anscombe's unmediated access comments seriously, but without positing a "distinctly conceived subject" with mysterious causal power that achieve such access—because such access is constitutive of Anscombean persons. Put metaphorically, the person provides a kind of "bridge" between the word 'I' and unmediated access: 'I' is connected to persons, as per our "deflated referent" story; persons, for Anscombe, are inherently connected to thoughts of actions and intentions; which thoughts are connected, for her, in an unmediated way to movements, postures, etc.

C. *Deflated Explanatory Burden*

We have argued that, for Anscombe, 'I' is associated with a deflated reference relation, and that the kind of things that 'I' in fact tends to specify, at a context, are deflated entities. But there is a third aspect to our deflationary approach. That aspect concerns issues about epistemology and explanation; and our inspiration this time is the work of Emma Borg (2004), and Herman Cappelen and Ernie Lepore (2004).

The general idea that we draw upon is that the lexical semantics for words should not be expected to explain, all on their own, their associated psychology, epistemology, and metaphysics. To give but one example, the semantics of 'rich' need not specify how much money a person must have in order to be rich. Still less must the semantics of 'rich' address the philosophical question of whether a society with extremely rich people and extremely poor ones can be just. In a similar vein, there will be aspects of the use of 'I', and of 'I'-users, that needn't be explained by the pronoun's "mode of meaning": think here of the peculiarities of self-knowledge, or the conditions for the persistence of persons over time.

Reading in deflation of this third sort is, we concede, a bigger exegetical stretch. There is solid textual evidence in "The First Person" for ascribing a mere rule-of-use which, as a matter of fact, applies to persons. In this case, the main motivation is different, driven more by read-the-text-as-promising considerations. As hinted, the three deflationary moves don't entail each other. Nonetheless, all play an essential role: in that sense, they require each other. Specifically, given deflation of the other two sorts, "explanatory deflation" is necessary to account for some of Anscombe's observations.

The exegetical stretch notwithstanding, there are some fit-with-the-corpus considerations that merit mention. First, it pays to remember that Anscombe's general philosophical methodology is reminiscent of J.L. Austin's: cautious not just in preaching but in practice; open to complexities and nuanced details; and comfortable with unresolved *aporias*. (*Intention* is an obvious, and brilliant, example.) In other work Anscombe at least sometimes approached philosophical problems with a divide-and-conquer attitude. To mention one especially notorious example, in her "Modern Moral Philosophy" (1958) she holds that there are some issues that are properly the burden of philosophy of psychology rather than of moral philosophy *per se*. Second, Anscombe was aware that phenomena of a similar nature arose in the absence of the lexical item 'I', e.g. in words such as 'now' and 'here'. Indeed, she mentions Casteñeda (1967) in a footnote. Similarly, she recognized that the same sort of phenomena show up with third person pronouns. For instance, 'Rob wanted to win, but didn't know this' is *not* made true by Rob wanting the 50 year-old Canadian philosopher to win, even though Rob was the 50 year-old Canadian philosopher in question. Anscombe was also keenly aware of first person *thoughts*, which arguably are not to be explained wholly by features of English pronouns. Third, there is one clear bit of textual support for our attribution of "explanatory deflation". Anscombe writes: "There is no objection to the topic of reidentification of selves—it is one of the main interests of the philosophers who write about selves—but this is not any part of the role of 'I'" (52–53).

It's worth ending this section by stressing, to avoid misunderstanding, that we are not discounting Anscombe's insights about the specialness of the first person. We take them very seriously. It is, however, consistent with that to expect that something beyond the context-sensitive rule-of-use for 'I' will help account for them.

5. *Defending our Answer*

This concludes our presentation of our proposed "Anscombean reference for 'I'". We turn more squarely now to the task of championing it. That requires defending it with respect to both our desiderata: the better our reconstruction fits with the text, the better; and the more promising the reconstruction is *qua* substantive account of the phenomena being investigated, the better.

Applied to any reconstruction of Anscombe's "The First Person", this yields two constraints: first, any putative reconstruction must comport with the four positive observations made by Anscombe; and second, any putative reconstruction must be at least initially promising and worthy of further investigation and development as a view about what we have been calling the "mode of meaning" of 'I'. We will defend our account first by addressing both constraints, and then by responding to some objections.

Let us begin with Anscombe's four positive observations about 'I' to see how our proposed reconstruction fits with them. Our proposed rule-of-use for 'I'—namely that if *X* makes assertions with 'I' as subject, then those assertions will be true if and only if the predicates used assertively are true of *X*—together with facts about what contexts of utterance almost always look like in our world, explains the near guarantee that any given use of 'I' will have a referent. Setting aside some famously puzzling cases (*cf.* Predelli 2005), there will almost always be a speaker in the context of utterance to serve as the target of the rule. Regarding misidentification, because no referent is intended with a pure indexical, there's no possibility of an error-inducing conflict between the intended referent and what the rule-of-use specifies.

Second, and again because of the associated rule-of-use for 'I', it follows that where 'I' is used there typically won't be genuine doubt that there is a speaker. Granted, full-blown Cartesian-style immunity to doubt isn't automatically ruled out by our reconstruction. Like Wittgenstein, however, Anscombe herself was very skeptical about claims of infallible first-person knowledge of facts. Moreover, embracing "deflated reference", it ceases to be a task of the semantic rule for 'I' to explain entirely on its own the perplexing epistemology of self-knowledge. To demand that is patently to demand too much.

Third, because there is no descriptive sense associated with 'I', there is no prediction that the output of the semantic rule associated with 'I' must be fixed via bodily properties, nor even via sensory ones. What if, *qua* metaphysician of mind, one wants every 'I'-user to be bodily? We have no problem with such a proposal. Indeed, as noted, our proposed rule-of-use is consistent with it. But our claim, again, is that you shouldn't ask the lexical entry for the first-person pronoun, all by itself, to guarantee that for you. The rule is "metaphysically silent" in that regard.

Fourth, and finally, consider in connection with the special action-guiding nature of 'I' two points that lead to the same result. David Kaplan (1989) and John Perry (1979) have proposed that the "character" for a word can play an autonomous role in generating action. If they are correct, we can already expect 'Rob Stainton is in danger' and 'I am in danger' to have different behavioral proclivities because of *how* the object gets specified when 'I', as opposed to a name, is used.[7] Second, as explained above, for Anscombe thoughts of actions, intentions, movements, etc., are unmediated and non-observational; and these features are central to persons in her forensic sense. So, given the deflated referent we are proposing, there will exist a special connection between what a token of 'I' specifies (hereabouts) and dispositions to act.

Reading Anscombe as working implicitly with a "deflated" notion of reference would fit well with her important observations about some philosophical peculiarities of 'I'. Our revisionist reading has another advantage: it doesn't commit Anscombe to glaringly false predictions about the syntactic and logico-semantic behavior of 'I'. To explain, we will first contrast the linguistic behavior of the expletive 'it' with that of noun phrases which exhibit a relatively "deflated" kind of reference. We then show that 'I' *obviously* patterns with the latter. This makes our reading the more charitable of the two.

In terms of syntax, being a "dummy element" with no reference, the expletive pronoun 'it' cannot license aphonic gaps which themselves have a referring role. Thus consider (1):

1. *It$_1$ seems that John is rich and [e$_1$ allowed him to buy the house]

This sentence strikes us as full-on ungrammatical; there's no question that it's odd. The reason is not that its meaning would be peculiar: 'It seems that John is rich and that fact allowed him to buy the house' is a perfectly fine way of expressing the thought which (1) gestures at. Instead, the issue is that the dummy subject 'it' is genuinely radically non-referring—so, the expletive provides no reference-source for the unpronounced subject of the second conjunct. (Consider also the strange-sounding 'It$_1$ fell to –20 degrees and e$_1$ froze the pipes'.) Being radically non-referring, the expletive 'it' also cannot form referential nominal compounds:

[7] In their discussions of "The First Person", both O'Brien (1994) and Rumfitt (1994: 625*ff*) make suggestions very roughly along these lines. O'Brien (1994: 280) suggests, e.g., that mastery of the rule-of-use for 'I' will *ipso facto* bring to light the metalinguistic fact that 'I' is a device of *reflexive* self-reference.

2. *[It and [the cloud]] seem likely to pour rain

And 'it' cannot receive focal stress; nor can it appear unembedded:

3. **It* seems likely to pour rain
4. Andrew: Do you expect snow?
   Rob: *It, it!'

Regarding logico-semantic features, the 'it' in question goes with zero-place predicates: the whole raison d'être of an expletive is to serve as surface subject to inflected verbs which do not take genuine arguments. Relatedly, sentences with expletive subjects do not license existential generalization. Witness the bizarreness of:

5. It is raining in Florida and it is snowing in Wisconsin. Therefore, there is something which is raining in Florida and snowing in Wisconsin

We now contrast how other "deflated referents" work, in terms of their syntax and logico-semantics. We will consider two examples: 'that rain storm' and 'his longstanding silence'.

As a preliminary, it's worth highlighting the respects in which these two count as "deflated" by our lights. In each case, the referent lies closer to the "feature-placing" end of the spectrum-of-individuals as opposed to its "self-standing subject" end. Relatedly, it is hard to individuate rain storms and long silences, and hard to count them. Turning from the referent to the reference relation, because of the context-sensitivity built into 'that' and 'his', in both examples the reference is not fixed solely by a descriptive sense. Finally, the explanatory powers of rain storms and longstanding silences are *comparatively impoverished*: e.g., it offers no great insight to explain precipitation by appeal to a rain storm, nor quiet by appeal to a long silence. (To explain the italics above: we do grant that, e.g., one can explain a puddle by appeal to a recent rain storm, and a baby's successful nap by appeal to silence around the house. It is the relative depth and nature of the explanation which is at issue: in these examples, they incline towards the "deflated".)

Now, such deflated noun phrases do license aphonic gaps which, in their turn, refer:

6. [That rain storm]$_1$ lasted for hours and e$_1$ was really frightening
7. Irma had a meeting with Ahmed. She called for [his longstanding silence]$_1$ [e$_1$ to end]

Both can serve as constituents in nominal compounds which themselves serve as referential-type arguments. For instance, in (8) 'that rain storm' conjoins with 'the dog which kept barking' to yield an argument to 'kept Sean awake'.

8. [$_{NP}$ [That rain storm] and [the dog which kept barking]] kept Sean awake

Example (9) illustrates the same point:

9. I am fed up with [$_{NP}$ [his ugly mug] and [his longstanding silence]]

As a final point about syntax, both 'that rain storm' and 'his longstanding silence' can receive stress and appear unembedded:

10. *That rain storm* kept me awake, not the barking dog
11. That rain storm! That damned rain storm!
12. I can live with Ahmed's messiness. But *his longstanding silence* drives me mad!

So much for their syntax. The logico-semantics of comparatively "deflated" noun phrases is also very different from expletives. They go with verbs that take genuine arguments, as illustrated already by (6), (10) and (12). And from these sentences one can draw valid existential inferences: respectively, that something lasted for hours; that something kept Andrew awake; and that something drives the speaker mad.

So much for the contrasts. As *we* interpret Anscombe, her view predicts that 'I' should pattern with 'that rain storm' and 'his longstanding silence', not with the 'it' of 'it seems' and 'it's raining'. This prediction is borne out.

'I' licenses anaphoric gaps and 'I' coordinates with patently referential nouns to form nominal compounds:

13. [I$_1$ want [e$_1$ to dance] or [e$_1$ to leave]]
14. [[$_{NP}$ [$_{NP}$ John] and I] love jazz]

The first person pronoun in English can readily receive stress, as in (15). And it can appear unembedded (in the accusative case), as in (16):

15. *I* won the race, not Ahmed
16. Andrew: Who wants tickets to Radiohead?
    Rob: Me, me!

Like clear cases of "deflated" noun phrases, the first person pronoun can also serve as argument to predicates generally, whatever their arity: 'I smoke', 'Alice likes me' and 'Alice gave me a book' are all perfectly fine. (Relatedly, if Irma says 'I smoke' and Ahmed says 'Irma smokes', they agree. Notice that this is not predicted by the 'I'-as-expletive view.) Finally, comparable to (6), (10) and (12), sentences containing 'I'/'me' license existential generalization: 'I smoke' entails that there exists something which smokes; 'Alice likes me' entails that there exists something which Alice likes, etc.

Our brief discussion of the syntax and logico-semantics of 'I' shows, on the one hand, that our reconstruction is promising as a substantive account of the first person pronoun's "mode of meaning". On the other hand, our reconstruction is exegetically superior because it avoids committing Anscombe to a range of obvious falsehoods about how 'I' behaves linguistically. This completes our positive defense of the reconstruction. We turn, in the next section, to objections that require rebuttal.

6. *Objections and Replies*

We have now canvassed what we take to be an initially plausible "variety of reference" (as one might nowadays call it) that can be extracted from Anscombe's "The First Person", and we have argued that our revisionist reading of "The First Person" meets our two desiderata: it complies with Anscombe's four positive observations, and it is independently promising as a view about what we have been calling the "mode of meaning" of 'I'. Moreover, *if* Anscombe was grasping for such an account of the semantic functioning (or "mode of meaning") of 'I', then, far from having committed an egregious linguistic blunder in that famous paper, she was in fact anticipating ideas that remain prevalent and important today.

But was she? Before defending 'Yes' as the appropriate answer, a reminder about our project is in order. If we were undertaking philosophy-focused history, a number of avenues of research would suggest themselves immediately. One could look into whether Anscombe's correspondence provides evidence of such a view, or whether marginal notes in the works she was reading at the time suggest it. One could try to trace which exact passages in her fellow Oxbridge philosophers might have inspired such a position on the linguistic role of 'I', etc.[8] Such questions—fascinating and worth pursuing—are not, however, our concern in the present paper. Still, a charitable and insightful reconstruction of the paper's arguments and conclusions requires, at a minimum, two things: first, internal consistency; and second, consistency with the philosophical milieu in which she was working and writing. So let us turn to some objections that touch on these considerations.

A. *On 'Referring'*

The first objection to our reading of "The First Person" is straightforward: our proposed interpretation simply doesn't fit with all the things Anscombe says about 'I' not referring. With this general observation we agree. But as argued in Stainton (2018), this complaint is merely terminological. Anscombe's (at that time perfectly apt) use of the vocable /rɛf(ə)r(ə)ns/ does not entail, even for Anscombe, that 'I' fails to have a rule-of-reference in *our 21st Century sense of 'reference'*. Our claim, recall, is that as we in 2018 might phrase it Anscombe is merely urging that in *one historically specific sense of 'refer'* that she was working with, 'I' doesn't "refer". But it is perfectly consistent with this view that 'I' can be used to "speak of" something; to "concern an object"; to "make an assertion about" something; and to "specify" an object". In other words, on our revisionist reading 'I' *does* refer for Anscombe, at least on

[8] There is also the concern that, so far as we have been able to establish, Anscombe never regretted nor retracted the phraseology of "The First Person" once "thinner" notions of reference became more standard.

the modern understanding of 'refer'. (Anscombe writes: "a self *can* be thought of as what 'I' stands for, or indicates, without taking 'I' as a proper name" (52). This seems to endorse the idea that 'I' does indeed "refer" in our "deflated" sense, but not in the "inflated" way that proper names were assumed to.)

B. *On "Missing Discussions"*

A second objection. There are discussions that one would expect to find in Anscombe's text if our "deflated" re-reading were correct—discussions, in particular, of other kinds of "deflated" referring which seem to belong in the same ballpark. Specifically, one would expect to find treatments of other terms which have only a rule-of-use that requires no intention, such as 'today' and 'here'. If she were offering a deflationary take on the "mode of meaning" of 'I', and if she really was concerned to put forward an allegedly novel variety of reference, surely she would have discussed similar context-sensitive words? They would be grist for her mill, if our interpretation were on the right track. And wouldn't she address Kripke-style views of names, according to which even they aren't "inflated"?

One can't explain away these seeming lacunae in terms of a lack of knowledge or a mere oversight. Anscombe was clearly aware of the existence of such context-sensitive items: again, she cites Castañeda (1967) in connection with the distinction between direct and indirect reflexive uses of 'I' and other pronouns. Similarly for names as directly referential: Anscombe mentions Kripke, in particular criticizing him for trying to recast the Cartesian argument in a way that downplays the centrality of 'I'.

Our reply has to do with the central aim of "The First Person". It is too seldom stressed that its objective is to rebut a neo-Cartesian semantic argument for mind-body dualism. That argument contains as a premise, in effect, that 'I' has "inflated" reference, and that this fixes the nominatum as non-bodily. That is, the very first paragraph of "The First Person" is not a mere historical preamble, but instead states the topic of the paper:

> Descartes and St. Augustine share not only the argument *Cogito ergo sum*—in Augustine *Si fallor, sum*—but also the corollary argument claiming to prove that *the mind* (Augustine) or, as Descartes puts it, *this I*, is not any kind of body... The first-person character of Descartes' argument means that each person must administer it to himself in the first person; and the assent to St Augustine's various propositions will equally be made, if at all, by appropriating them in the first person. In these writers there is the assumption that when one says 'I' or 'the mind', one is *naming something* such that the knowledge of its existence, which is a knowledge of itself as thinking in all the various modes, determines what it is that is known to exist (45, our emphasis).

Given this focus on 'I' as name-like, there is a good reason why Anscombe would by-pass the workings of 'here', 'today', etc. Her focus was rightly on how 'I' *does not* work. *We* are proposing a positive account of 'I''s functioning, based on clues from the text. But Anscombe's aim was different: it was to shut down this neo-Cartesian argument at its very outset. In light of this, though perhaps it would have been illuminating as an aside, a discussion of other terms in the same "deflated" ballpark would have been just that: an aside. (Anscombe writes: "To say all this is to treat 'I' as a sort of proper name. That's what gets us into this jam" (48).)

As for why she elided discussion of Kripke's views, a first point is that early 70s Oxbridge had simply not yet embraced his lessons about direct reference. In any case, Anscombe just *does* take proper names to be sense-bearing; this seems to be non-negotiable for her. What's more, if Kripke were right that even names lacked descriptive senses, then the neo-Cartesian semantic argument couldn't get off the ground. Thus Kripke, far from proving an opponent, would be offering up another path to the same no-sense-for-'I' conclusion.

Thus, whatever she may have had in mind as she wrote, it was perfectly reasonable for Anscombe to have avoided making positive claims about semantic similarities between 'I' on the one hand and other "pure indexicals" on the other. And it was perfectly reasonable for her to sidestep discussion of Kripke's newfangled views on names.

C. *On Identity Propositions*

A third objection is arguably the most pressing. Recall that according to our revisionist interpretation of "The First Person" the first-person pronoun 'I' *is* associated with a referent. But if that's the case, how could Anscombe hold that a sentence of the form 'I am Elizabeth Anscombe' does not express an identity proposition? Worse, her infamous claim fits very well with the "Straight" reading that we are challenging: if 'I' is an expletive, then of course 'I am Elizabeth Anscombe' will not express an identity.

We have three replies. First, because the referent is deflated on the view we are attributing to Anscombe, there isn't *the right kind of object* for an identity. Being merely a locus for feature-placing, there are no clear individuation conditions for the thing "spoken of"/"specified" by 'I'; so if genuine identity requires "robust" objects satisfying precise Quinean individuation conditions, then it follows that there won't be person-involving identities (in that exigent sense).

Second, because the reference relation is deflated, there are not two senses, each corresponding to the same object. But since Fregeans require this for an (informative) identity statement, there won't by Anscombe's lights be any such statements involving 'I'.

Third, because of the deflated "metaphysical and explanatory power" of 'I', the first person pronoun on its own does not fix or entail the nature of the thing-which-asserts. In particular, it does not fix it as a kind of body. In contrast, according to Anscombe the proper name 'Elizabeth Anscombe', as a matter of analytic entailment, must refer to a certain kind of animal, namely a human, female animal. Worse for "real identities", and revisiting a point from footnote 4, Anscombe's view seems to be that 'I'-users turn out to be persons in the forensic sense, and these are only "intimately connected" with bodies. So again, 'I am Elizabeth Anscombe' cannot state an identity proposition in the relevant sense. (That Anscombe is rejecting only "identity statements" *construed in some philosophically strict way* is suggested by her acknowledgement at the outset of her paper that there is a "mundane, practical, everyday sense" in which 'I am Descartes' can be true (46).)

Here is another way at our main point. Ask: why, according to Anscombe, is 'Elizabeth is Anscombe' a genuine identity statement? The answer is: because the referent of both 'Elizabeth' and 'Anscombe' is a robust, countable human body; and because there is a sense associated with both proper names, each yielding the same nominatum. Also, we have not just an intimate connection between Elizabeth and Anscombe, but one single thing. Now compare this with the case of 'I'.

7. *Conclusion*

Many readers have taken Anscombe to hold a radical non-referring view about 'I', according to which 'I' is a sort of expletive pronoun. Such a view, however, fits poorly with numerous points made explicitly by Anscombe in her paper; it is also manifestly incorrect about both the surface syntax and logico-semantics of 'I'. Fair engagement both with Anscombe as a founder of the Analytic tradition and with her exceptionally insightful paper requires us, therefore, to identify a "mode of meaning" for 'I' that coheres better with her text and with her larger philosophy, as well as with certain empirically obvious facts about the first person pronoun.

With that in mind we have proposed an "Anscombean reference for 'I'" which is deflated along three axes: first, the reference *relation* does not involve a descriptive sense, but only a rule-of-use where intentions are otiose; second, the *referent* is a "person" in the forensic sense of that term; and third, the *explanatory burden* of "Anscombean reference" in epistemology, psychology, and metaphysics is fairly limited, so that many of the puzzling aspects of the first person must be explained by something other than the lexical semantics of 'I'.

References

Anscombe, E. 1957. *Intention.* Oxford: Blackwell.

_____1958. "Modern Moral Philosophy." *Philosophy* 33 (124): 1–19.

_____1975. "The First Person." In S. Guttenplan (ed.). *Mind and Language.* Oxford: Oxford University Press: 45–65.

Austin, J. L. 1957. "A Plea for Excuses." Reprinted in J. O. Urmson and G. J. Warnock (eds.) 1979. *Philosophical Papers.* Oxford: Oxford University Press: 174–204.

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

Cappelen, H. and E. Lepore. 2004. *Insensitive Semantics.* Oxford: Blackwell.

Castañeda, H.-N. 1967. "The Logic of Self-Knowledge." *Nous* 1: 9–22.

Clarke, D. S. 1978. "The Addressing Function of 'I.'" *Analysis* 38 (2): 91–93.

Evans, G. 1982. *Varieties of Reference.* Oxford: Oxford University Press.

Garrett, B. J. 1994. "Anscombe and the First Person." *Crítica* 26 (78): 97–113.

_____1997. "Anscombe on 'I.'" *Philosophical Quarterly* 47 (189): 507–511.

Hamilton, A. 1991. "Anscombean and Cartesian Scepticism." *The Philosophical Quarterly* 41 (162): 39–54.

Hinton, E. 2008. "Anscombe's First Person." *Prometheus Journal.* Accessed at http://prometheus-journal.com/2008/12/23/anscombe%E2%80%99s-first-person/.

Kaplan, D. 1989. "Demonstratives." In J. Almog, J. Perry and H. Wettstein (eds.). *Themes From Kaplan.* Oxford: Oxford University Press: 481–563.

Kripke, S. 1980. *Naming and Necessity.* Cambridge: Harvard University Press.

_____2011. "The First Person." In his *Philosophical Troubles.* Oxford: Oxford University Press: 292–321.

O'Brien, L. 1994. "Anscombe and the Self-Reference Rule." *Analysis* 54 (4): 277–281.

Peacock, C. 2008. *Truly Understood.* Oxford: Oxford University Press.

Perry, J. 1979. "The Problem of the Essential Indexical." *Nous* 13 (1): 3–21.

Predelli, S. 2005. *Contexts.* Oxford: Oxford University Press.

Quine, W. V. O. 1948. "On What There Is." *Review of Metaphysics* 2 (1): 21–38.

Rumfitt, I. 1994. "Frege's Theory of Predication." *The Philosophical Review* 103 (4): 599–637.

Stainton, R. J. 2018. "Re-Reading Anscombe on 'I.'" *Canadian Journal of Philosophy* 49 (1): 70–93.

Strawson, P. F. 1953. "Particular and General". *Proceedings of the Aristotelian Society* 54: 233–260.

_____1959. *Individuals.* London: Routledge.

_____1966. "Self, Mind and Body." *Common Factor* 4: 5–13.

Taschek, W. 1985. "Referring to Oneself." *Canadian Journal of Philosophy* 15 (4): 629–652.

Teichmann, R. 2008. *The Philosophy of Elizabeth Anscombe.* Oxford: Oxford University Press.

Thomasson, A. 2010. *Ordinary Objects.* Oxford: Oxford University Press.

van Inwagen, P. 2001. "'I am Elizabeth Anscombe' is not an Identity Proposition." *Metaphysica* 2 (1): 5–8.

Wiseman, R. 2017. "What Am I and What Am I Doing?" *The Journal of Philosophy* 114 (10): 536–550.

Croatian Journal of Philosophy
Vol. XVIII, No. 54, 2018

Negative or Positive? Three Theories of Evaluation Reversal

BIANCA CEPOLLARO
San Raffaele University, Milan, Italy

In this paper, I consider the phenomenon of evaluation reversal for two classes of evaluative terms that have received a great deal of attention in philosophy of language and linguistics: slurs and thick terms. I consider three approaches to analyze evaluation reversal: (i) lexical deflationist account, (ii) ambiguity account and (iii) echoic account. My purpose is mostly negative: my aim is to underline the shortcomings of these three strategies, in order to possibly pave the way for more suitable accounts.

Keywords: Slurs, thick terms, reclamation, evaluation reversal, ambiguity, echo.

1. *Introduction*

Language is not only used to describe state of affairs, but also to evaluate them, i.e., to express subjective judgements. The most prototypical pieces of language that are employed for the purpose of evaluating are thin terms, such as 'good' and 'bad', but many other expressions systematically convey evaluative contents: just to mention a few, the so-called thick terms, slurs, aesthetic predicates, predicates of personal taste and the like.

In this paper, I assess the phenomenon of evaluation reversal: uses of language in which a term that typically carries an evaluative content with a certain polarity (positive or negative) can be felicitously used in order to convey evaluative content with an *opposite* polarity (from positive to negative and vice versa). In this work, I focus on slurs and thick terms—expressions which are *systematically* associated with evaluative contents—, while I leave aside descriptive terms that can be *on occasion* used evaluatively (see Stojanovic 2016a, especially section 2.1).

Slurs are derogatory terms targeting individuals or groups on the basis of their belonging to a certain category.[1] Prototypical English slurs target nationality, ethnic origins, sexual orientation, religion and so on, and they are associated with a negative evaluative content. Thick terms, on the other hand, are usually defined as those expressions which combine descriptive and evaluative contents, both positive and negative:[2] 'generous' for instance does not only refer to the property of being willing to share one's resources, but it also conveys the idea that it is good to be so; 'lewd' refers to the property of being sexually explicit beyond conventional boundaries, but it also conveys the idea that it is bad to be so. In this work, I do not go through all the possible theories of slurs and thick terms; instead I focus on the case of evaluation reversal and critically discuss three accounts.

The paper goes as follows. In section 2 I briefly present two phenomena which can be accounted for in terms of evaluation reversal: the reclamation of slurs and the variability of thick terms. In section 3, I discuss three theories developed to account for reclamation or variability (or both): they are the lexical deflationist account (3.1), the ambiguity account (3.2) and the echoic account (3.3). My goal is to pinpoint the shortcomings of each of them. My aim here is strictly negative, but clarifying the difficulties of each approach should pave the way for more promising accounts.

2. *Evaluation reversal: reclamation and variability*

This section is dedicated to the reclamation of slurs (section 2.1) and variability of thick terms (section 2.2). In this paper, I do not develop an argument to support the thesis that the two phenomena are similar under crucial aspects (for a defense of a similar position, see Cepollaro 2017a), but I do treat both of them as cases where a lexical item conventionally associated with a positive or negative evaluation can be used on occasion with the opposite polarity.

2.1 *The reclamation of slurs*

In the debate on slurs, scholars underline how these expressions systematically convey derogatory contents towards the target group regardless of (i) how the slur is embedded and (ii) what the intentions of the speaker are. As for (i), we observe that an utterance like 'Lea is a wop' keeps being derogatory also when it is embedded under negation, conditional, modal, question: 'Lea is not wop', 'If Lea is a wop, her son is too', 'Lea may be a wop', 'Is Lea a wop?'. The relation between slurs and derogation is such that the pejorative content resists when embedded

[1] See i.a. Potts (2005), Hom (2008), Anderson and Lepore (2013a, 2013b), Camp (2013), Cepollaro (2015), Jeshion (2013), Bolinger (2017).

[2] See Hare (1963), Williams (1985), Blackburn (1992), Gibbard (1992), among others.

under semantic operators. As for (ii), consider a case where someone calls a person a slur and then apologizes by saying she did not mean to offend. The absence of the *intention* to offend is not enough to cancel or neutralize the derogation:[3] slurs *are* demeaning, notwithstanding the intentions of the speaker.

However, we should not take these observations as evidence that there is no way in which slurs occur without being derogatory. As a matter of fact, slurs can also display some peculiar uses, that go under the label of 'reclamation', which seem to convey no derogation. Reclamation is the phenomenon for which the members of a target group can use the slur targeting their own group in such a way that slurs are not derogatory in those cases. Reclamation constitutes a challenge to a theory of slurs which aims to account for the fact that the pejorative content of these expressions seems to resist all kinds of embedding and attempts of neutralization. The phenomenon raises many questions, some of which we will discuss here. Among the main issues scholars are faced with there is the question as to whether reclaimed uses of slurs are literal uses of language; as to whether, once a slur gets reclaimed, it is still the same lexical item as before; as to whether reclaimed uses of slurs pose similar moral problems as non-reclaimed ones; as to whether reclamation can take place without political awareness or not, and so on.

To complicate the picture even more, as Jeshion (ms) underlines, reclamation is not a uniform and homogeneous phenomenon: there are many ways in which a slur can be used by in-groups without being derogatory. Some reclaimed uses of slurs convey positive evaluative content, some are just non-negative without being necessarily positive; some are possible for in-groups only, while some are available for out-groups too; some sound ironic, satirical or sarcastic, while some do not, and so on. In this work, I am interested in reclaimed uses of slurs where the term is used in a positive way, that is, in the cases where the evaluation conveyed by these expressions is *reversed*, not just suspended (for an analysis of evaluation suspension, see Cepollaro 2017a: section 3.2). It may turn out that this is just a subgroup of reclamation in general.

2.2 *The variability of thick terms*

Scholars in ethics and metaethics noticed that even though thick terms are associated with evaluative contents linked with a certain polar-

[3] In the last decade, quite a few of these cases made it to the newspaper. What they all have in common is that someone used a slur and then tried to apologize by appealing to their own non-derogatory intentions; in all of those cases, this attempted apology failed to excuse them, as in the case of slurs the absence of a derogatory intention does not typically cancel the derogation which did take place nevertheless. Just to mention three such cases: https://www.theguardian.com/sport/2015/dec/15/rajon-rondo-gay-slur-nba; https://www.washingtonpost.com/news/early-lead/wp/2017/10/31/conor-mcgregor-apologizes-for-homophobic-slur/?utm_term=.d78b2c9128fa; https://www.huffingtonpost.com/entry/nhl-athlete-non-apology_us_5922fcace4b094cdba55ecb0?guccounter=1.

ity, on occasion they can be felicitously used with an opposite polarity.[4] Just to make an example, provided that 'chaste' typically carries a positive evaluation, under certain circumstances, it can be used in a negative way. This example is taken from the Corpus Of Contemporary American-English (or COCA; Davies 2008):

> "Not sure how long I'll be gone. (...)'. Elaine gave him a quick kiss on the cheek. 'That was a little chaste'. 'Don't look now, but we seem to be of interest to about fifty elderly women on the tour bus behind you' 'Should we give them something to stare at?"

It looks like the speaker is using 'chaste' as a negative thing for a kiss to be; despite the use of an evaluative term with a positive polarity, the speaker is *not* endorsing that kind of positive evaluation, quite the contrary: he is using 'chaste' as meaning to convey a negative rather than positive evaluation. The example which is mostly discussed in the literature is the positive use of a thick term with negative polarity, namely 'lewd'. The original example from Blackburn is the following:

> "[We may] worry that this year's Carnival was not lewd enough" (Blackburn 1992: 296, quoted in Väyrynen 2013: 217).

But Väyrynen (2013) changes the example a bit in order to avoid the unneeded complications brought about by the expression 'not enough' and credits Matti Eklund for the final version of the example:

> "The carnival was a lot of fun. But something was missing. It just wasn't lewd. I hope it'll be lewd next year" (Väyrynen 2013: 85)

The speaker is using 'lewd', typically associated with a negative evaluation, as expressing a positive one. As in the case of reclamation, scholars need to address questions e.g. whether instances of variability count as literal uses of thick terms or as to whether, once an expression like 'lewd' gets used positively, it is still an instance of the same lexical item as before.

3. *Three theories of evaluation reversal: lexical deflationism, ambiguity, echo*

In this section, I consider three possible approaches to evaluation reversal and I apply them to the case of reclamation and variability. These strategies have been explicitly proposed to account for the phenomenon of evaluation reversal in relation to thick terms specifically (this is the case for the lexical deflationist account, section 3.1), or to slurs (this is the case for ambiguity account, section 3.2) or to both slurs and thick terms (echoic account, section 3.3). In what follows, I try to see how each of these approaches can explain evaluation reversal for both slurs and thick terms. As announced, my aim is negative: my goal is to underline the shortcomings of the three strategies in order to pave the way for more promising accounts.

[4] Hare (1952), Blackburn (1992), Väyrynen (2011, 2013), Eklund (2013).

3.1 *Lexical deflationism*

The lexical deflationist account of evaluation reversal was put forward for the variability of thick terms rather than for the reclamation of slurs. However I assess its plausibility both for slurs and thick terms, by following a suggestion of Väyrynen (2016).

Lexical deflationism amounts to the idea that the reason why the evaluation conveyed by slurs and thick terms can change polarity on occasion is that it is not lexically encoded. For this approach, the evaluative content with which these expressions are associated consists in pragmatic implications; the addressees infer the evaluative content (and of course its polarity) in each context. Väyrynen (2013) defends a similar thesis for thick terms; moreover, Väyrynen (2016) hints at the possibility to develop a theory of slurs along similar lines; to him, the resulting approach would resemble Bolinger (2017)'s one. For Bolinger, the derogatory content of slurs is due to purely pragmatic mechanisms:

> In choosing to use a slurring term rather than its neutral counterpart, the speaker signals that she endorses the term (and its associations). Such an endorsement warrants offense, and consequently slurs generate offense whenever a speaker's use demonstrates a contrastive preference for the slurring term. (Bolinger 2017: 439)

In this framework, when speakers reclaim a slur, they use it defiantly, without endorsing the relevant associations; as the group of speakers who do so grows, the link between the lexical item and the associated contents grows weaker and weaker. When reclamation reaches a certain stage, it is the context that determines each time whether the slur carries a negative evaluative content or not. Such a strategy, defended by Väyrynen and arguably by Bolinger too, appears to analyze slurs and thick terms in a way that makes them similar to terms that do not lexically encode evaluation, but can be used in evaluatively on occasion—either positively or negatively—, for instance 'intense' (see Stojanovic 2016a, 2016b about "valence-underspecification").

Let us now look at the shortcomings of lexical deflationism. As far as thick terms are concerned, one may wonder if the context is really enough to determine the polarity of the evaluative content. In what follows, I propose a case which suggests, *contra* lexical deflationism, that it is the lexical content which determines the interpretation of the evaluation. Suppose there are two thick terms which share the same descriptive content such that one is typically associated with a positive evaluation and the other with a negative one. For the sake of the example, suppose that this is the case for 'reckless' and 'brave', so assume that their descriptive meaning amounts to something like 'willing to do something dangerous', while their evaluative contents have opposite polarities, one negative, one positive. Now suppose that two such terms are used in the same context:

A. What she did was courageous!
B. It was not. It was reckless.

The two speakers, A and B, agree on what the facts are and what the act at stake is and, nevertheless, disagree on how to evaluate it. If lexical deflationism were right, the audience of such a dialogue would be confused about how the two speakers evaluate the act at stake: since for this approach it is the context and not the lexical content which determines the interpretation of the evaluative content, if the context is one and the same and—by hypothesis—the descriptive content is the same, the context should attribute the same evaluative content to both of them. However, the audience of the dialogue has no difficulty in understanding that A approves of the action under discussion and B disapproves of it. I argue that this is so because the evaluation *is* in fact lexically encoded: competent speakers can come up with a default interpretation roughly corresponds to the conventional meaning of the term at stake.

As far as slurs are concerned, on the other hand, lexical deflationism in the version of Bolinger (2017) has some problems in accounting for the intuition that slurs are derogatory also in a context where they are speakers' default choice (e.g. racist environments and discussions). Let me state again that for lexical deflationists, slurs do not lexically encode offensive contents, they are only pragmatically associated with them as a matter of contrastive choice. If that was the case, then they would not be associated with any such content in a situation where they are the default choice. In other words, lexical deflationism can account for the intuition that slurs are derogatory (i.e. they convey offensive contents) when they occur in non-racist environments, but not when they occur in bigot contexts, where they do not trigger any pragmatic implication in virtue of being the default option. I take this as evidence that lexical deflationism is wrong in postulating that slurs do not convey derogatory contents at the level of conventional meaning.

To sum up, we started from observing that lexical deflationism has an easy way to explain evaluation reversal: since the evaluative content is *not* lexically encoded in the conventional meaning of slurs and thick terms, its polarity can change on occasion. However, we have observed that this approach has problems on its own explaining the behavior of slurs and thick terms *in general* and thus it may not be a viable option to account for evaluation reversal. In what follows, we consider two alternative theories which endorse the claim that slurs and thick terms lexically encode evaluative contents. The challenge which these approaches need to meet is to account for the possibilities for such evaluative contents to change polarity.

3.2 *Ambiguity*

The ambiguity account of evaluation reversal was put forward for the reclamation of slurs rather than for the variability of thick terms. However, as I did for the lexical deflationist approach, I shall consider both applications.

The main point of the ambiguity account of reclamation is that once slurs are reclaimed, a new word comes to exist. Many scholars endorsed this thesis[5] which it gets rid of the problem raised by reclamation at its source: it rejects the idea that a lexical item undergoes an evaluation reversal, since there are in fact two different lexical items. According to this approach, reclamation does not challenge those theories of slurs which analyze the derogatory content as part of the conventional meaning of the term, because in this framework reclaimed slurs are not instances of the same lexical items as slurs; in fact, they are not 'slurs' properly speaking. They are other terms with a different—and non-derogatory—meaning. The same would hold for thick terms: once a thick term is used with a different polarity, a new evaluative term comes to existence.

In the debate on slurs, this proposal has received some criticism from Anderson and Lepore (2013a) and Anderson (2018), an objection which Ritchie (2017) calls 'Reclamation Worry' (RW). The criticism is the following: if there was an ambiguity relation between reclaimed and non-reclaimed slurs, then any speaker would be able to felicitously use one or the other lexical item; however, this is famously not the case, as usually only in-groups and not out-groups can felicitously use the non-derogatory term. Anderson and Lepore use this argument against the theories of slurs which (i) are content-based (i.e. hold that these expressions lexically encode pejorative contents) and (ii) explain reclamation by relying on an ambiguity account. For Anderson and Lepore, because not every speaker can use any meaning of a slur (derogatory and non-derogatory), then the ambiguity thesis of reclamation must be wrong and therefore content-based theories should be rejected because they would have no other way to explain reclamation.[6]

In what follows, I present challenges to the ambiguity thesis that are orthogonal to the Reclamation Worry, as I do not take it to constitute a problem for the ambiguity thesis. In fact, in Cepollaro (2017b) I argued that, on a closer inspection, the RW should not trouble the defendants of the ambiguity account too much, because there are further cases in other languages (e.g. personal pronouns in French, German, Italian, Spanish) where two lexical items are ambiguous and the issue of which speaker can use which term is a matter of socially-determined factors. Leaving that worry aside, the main problem with the ambiguity thesis is that it raises more difficulties than it would have initially appeared. In particular, it needs a detailed characterization of 'ambiguity', which is something that scholars tended to overlook. As Anderson (2018) underlines: "Positing a lexical ambiguity, for example, would mean that

[5] Hom (2008: 428, 438), Richard (2008: 16), Saka (2007: 146–147), Miščević (2011: 176), Jeshion (2013: 250–253), Whiting (2013: 370).

[6] Section 3.3 shows that this is not the case: content-based views are also compatible with the echoic account; so, Anderson and Lepore's criticism would not suffice anyway to challenge content-based approaches, even if the objection towards the ambiguity thesis were correct.

either [the N-word] corresponds to non-identical entries in the lexicon or it expresses multiple meanings”. As Anderson remarks, there are two options available for the ambiguity account, which we can attribute to the phenomena of homonymy and polysemy. The first characterization—homonymy—boils down to analyze a standard and a reclaimed slur as corresponding to two different entries in the dictionary.[7] The second characterization—polysemy—is to posit that a standard and a reclaimed slur correspond to one lexical entry with multiple meanings.

Let me start from homonymy, which is the phenomenon for which two lexical items are written and pronounced in the same way and such a thing is—so to speak—accidental: there is no special connection (for instance at the level of etymology) between the two terms. This is the case for example for ‘bank’: we can talk about two different lexical entries bank_1 and bank_2, where the former refers to the financial institution and the latter to the river side. We can observe that the two items have different etymologies and that the two meanings corresponding to bank_1 and bank_2 are expressed by different words in other languages (‘banca’ vs. ‘sponda’ in Italian, ‘banque’ vs. ‘rive’ in French, etc.). If we look at the relation between a standard offensive use of a slur and a reclaimed one, we observe that it does not resemble homonymy: the two uses do not correspond to terms with different etymologies and the link between the offensive and the non-offensive use of the term does not amount to an accident, as in the case of bank_1 and bank_2.

The second option for the ambiguity thesis to characterize the relation between standard and reclaimed slurs is polysemy, the phenomenon for which one term has multiple meanings that correspond to different aspects. For instance, take ‘bottle’. The lexical item can refer to the object or to the content of the object, as in “The bin is full of empty bottles” (object) and “She drank two bottles of Pastis” (content). The two meanings—object and content—correspond to two related aspects of the concept BOTTLE. If we go back to slurs, we see that if ambiguity is characterized in terms of polysemy, standard and reclaimed slurs would have to correspond to different meanings of the same word. This sounds more promising than holding that the two are not related and that the ambiguity is merely accidental, as in the case of homonymy (see ‘bank’). However, the two meanings do not seem to correspond to two aspects of the same concept, as in prototypical cases of polysemy. If we look at instances of regular polysemy, we cannot really trace cases where the two aspects involved only differ at the level of evaluative rather than descriptive content. The same observations can be made for thick terms.

The ambiguity account needs deeper investigation on homonymy and polysemy in order to develop a detailed and precise proposal of

[7] Which is something Ritchie (2017) has in mind when she formulates the Reclamation Worry by noticing that “Anyone can use ‘bank’ to mean financial institution or side of a river”.

evaluation reversal, because, as it stands, there are too many dissimilarities between the homonymy and polysemy involving descriptive meanings on the one hand (see 'bank', 'bottle') and the case of evaluation reversal involving evaluative meanings on the other hand (see slurs and thick terms).

Finally, the ambiguity account needs a supplementary story about how the evaluation reversal begins in the first place: we know that for new lexical items (or new meanings of old words) to come to exist, certain conditions have to be met: the fact that a term is on occasion used with a different polarity than usual does not seem enough to postulate the creation of a new lexical item. The echoic theory that we discuss in the following section seems to be better equipped to account for how the reversal begins.

3.3 *Echo*

The echoic account of evaluation reversal was originally put forward for the reclamation of slurs by Bianchi (2014), furtherly supported by Miščević and Perhat (2016), and extended to the variability of thick terms in Cepollaro (2017a).

The bulk of the proposal is that the cases of evaluation reversal are instances of dissociative echoic uses of language, i.e. cases in which by uttering an evaluative locution the speaker is evoking the evaluative content conveyed by that particular term, but at the same time she is expressing her dissociation with respect to such content. Instances of evaluation reversal are not literal uses of evaluative language. As a matter of fact, the echoic theory was put forward by Sperber and Wilson (1986) in order to account for irony: in ironic utterances, speakers evoke some thought, belief or expectation that they attribute to someone else and at the same time they express their dissociation with respect to the evoked content. In this sense, evaluation reversal counts as a case of irony. Since irony involves a dissociative attitude, the possibility for irony to be successful (i.e. to be felicitous and to get recognized) requires a correct interpretation of attitudes. As a consequence, for evaluation reversal to be successful, the audience needs to recognize and correctly interpret the attitude of the speaker, which leaves room for all sorts of misunderstanding. Recall the example we mentioned in section 2.2., when the speaker complained about a kiss by saying "That was a little chaste". For the echoic theory, the speaker is evoking the evaluative content associated with 'chaste', namely 'it is good to be abstaining from sexual intercourse', and he is making fun of it by expressing his dissociative attitude. The same goes for slurs: when the actress, singer and stand-up-comedian Lea DeLaria calls herself 'that fucking dyke'; what she does is evoking the pejorative content associated with the homophobic slur and expressing her dissociation from it at the same time.

The echoic approach can tell a plausible story about how evaluation reversal starts: it starts by defiantly subverting the lexically encoded evaluation of a certain locution by means of irony. However, there are a few points on which the theory shows its weaknesses and call for adjustment.

Most of the difficulties that the approach has concern slurs rather than thick terms, for which, on the contrary, it seems to work quite well (for a contrary opinion, see Väyrynen 2013). The main issue is whether the echoic theory can account for all cases of reclamation. As noticed in section 2.1, reclamation is far from being a uniform and homogenous phenomenon: different instances display different properties. In particular, there are cases which are convincingly captured by the ironic explanation (for instance, the above-mentioned examples of uses of 'chaste' and 'dyke'). To support the view, notice that the clearer the ironic intentions are, the easier it is for the audience to understand that the usual evaluation is subverted. On the other hand, however, not *all* instances of reclamation appear to be ironic, not even in the technical sense which Sperber and Wilson have in mind. In particular, advanced-stage cases of reclamation seem to have lost the ironic flavor. Consider for example certain uses of 'queer': if one talks about the 'queer studies' class she is taking, it is implausible to postulate an ironic use of 'queer', it is just how the class is called; if one appreciates 'queer tango nights', there is no reason to imagine that she is being ironical, it is just how certain kinds of tango are called. In other words, when the process of reclamation is at an enough-advanced stage—i.e. when there is an attested non-derogatory use of the expression which used to be a slur—, the reclaimed uses can cease to sound ironic. Note that this feature (the absence of ironic flavor) does not depend on the fact that reclaimed uses of 'queer' become available for out-groups too: as a matter of fact, also some reclaimed uses of the 'n-word' which are available for in-groups only fail to display irony.

The fact that the echoic approach does not seem to account for *all* instances of reclamation can be taken to suggest either that reclamation is not a uniform and homogeneous phenomenon and that therefore new explanations are required—as Jeshion (ms) claims—or that the echoic account is well-equipped to account for some cases of reclamation but needs some sort of supplementary story for the non-ironic cases.

4. *Conclusion*

As stated at the beginning, this paper has a negative purpose, i.e. underling the shortcomings of three existing accounts of evaluation reversal. The analysis focused on two different cases of evaluatives—slurs and thick terms—in order to look at evaluation reversal with a broader stance. In particular, after presenting the phenomenon at stake (section 2), I argued that lexical deflationism has troubles explaining the behavior of slurs and thick terms in the first place and thus it should not be taken as a viable explanation of evaluation reversal (section 3.1); as for the ambiguity thesis, I showed that it lacks a detailed account of how standard cases of homonymy and polysemy relate to the case of evaluatives (section 3.2). Finally, I moved to the echoic approach (section 3.3) and underlined that despite its many merits, it displays some weaknesses in accounting for what appear to be non-ironical uses of reclaimed slurs. I hope that by clarifying the difficulties of each approach, I paved the way for more promising theories.

References

Anderson, L. and Lepore, E. 2013a. “Slurring words.” *Noûs* 47 (1): 25–48.

Anderson, L. and Lepore, E. 2013b. “What did you call me? Slurs as prohibited words.” *Analytic Philosophy* 54 (3): 350–363.

Bianchi, C. 2014. “Slurs and appropriation: An echoic account.” *Journal of Pragmatics* 66: 35–44.

Blackburn, S. 1992. “Through Thick and Thin.” *Proceedings of the Aristotelian Society*, Supplementary Volume 66: 284–299.

Bolinger, R. J. 2017. “The Pragmatics of Slurs.” *Noûs* 51 (3): 439–462.

Cepollaro, B. 2015. “In Defense of A Presuppositional Account Of Slurs.” *Language Sciences* 52: 36–45.

Cepollaro, B. 2017a. “When evaluation changes. An echoic account of appropriation and variability.” *Journal of Pragmatics* 117: 29–40.

Cepollaro, B. 2017b. “Let’s not worry about the Reclamation Worry.” *Croatian Journal of Philosophy* 17 (2): 181–194.

Davies, M. 2008. *The Corpus of Contemporary American English* (COCA Available online at https://corpus.byu.edu/coca/.

Eklund, M. 2013. “Evaluative Language and Evaluative Reality.” In Kirchin, S. (ed.). *Thick Concepts*. Oxford: Oxford University Press: 161–181.

Gibbard, A. 1992. “Thick Concepts and Warrant for Feelings.” *Proceedings of the Aristotelian Society*. Supplementary Volume 61: 267–283.

Hare, R. M. 1952. *The Language of Morals*. Oxford: Oxford University Press.

Hom, C. 2008. “The semantics of racial epithets.” *Journal of Philosophy* 105: 416–440.

Jeshion, R. 2013. “Expressivism and the Offensiveness of Slurs.” *Philosophical Perspectives* 27 (1): 231–259.

Jeshion, R. 2017. *Pride and Prejudiced*. Handout for the Evaluatives and Expressives Workshop (Milan).

Miščević, N. 2011. “Slurs & Thick Concepts. Is the New Expressivism Tenable?” *Croatian Journal of Philosophy* 11 (2): 159–182.

Miščević, N. and Perhat, J. 2016. *A word which bears a sword*. Zagreb: KruZak.

Potts, C. 2005. *The logic of conventional implicatures*. Oxford: Oxford University Press.

Richard, M. 2008. *When Truth Gives Out*. Oxford: Oxford University Press.

Ritchie, K. 2017. “Social Identity, Indexicality, and the Appropriation of Slurs.” *Croatian Journal of Philosophy* 17 (2): 155–180

Saka, P. 2007. *How to Think about Meaning*. Dordrecht: Springer.

Sperber, D. and Wilson, D. 1986. *Relevance: Communication and Cognition*. Cambridge: Harvard University Press.

Stojanovic, I. 2016a. "Evaluative adjectives and evaluative uses of ordinary adjectives." *Proceedings of LENLS12: Language Engineering and Natural Language Semantics*. The Japan Society for Artificial Intelligence: 138–150.

Stojanovic, I. 2016b. "Expressing aesthetic judgments in context." *Inquiry* 59 (6): 663–685.

Väyrynen, P. 2011. "Thick Concepts and Variability." *Philosophers' Imprint* 11: 1–17.

Väyrynen, P. 2013. *The lewd, the rude and the nasty: A study of thick concepts in ethics*. Oxford: Oxford University Press.

Väyrynen, P. 2014. "Essential Contestability and Evaluation." *Australasian Journal of Philosophy* 92 (3): 471–488.

Väyrynen, P. 2016a *Evaluatives and Pejoratives*. Handout for Linguistics Seminars-Scuola Normale Superiore, Pisa.

Whiting, D. 2013. "It's Not What You Said, It's the Way You Said It: Slurs and Conventional Implicatures." *Analytic Philosophy* 54 (3): 364–377.

Croatian Journal of Philosophy
Vol. XVIII, No. 54, 2018

Lewisian Scorekeeping and the Future

DEREK BALL
University of St Andrews, St Andrews, Scotland, UK

The purpose of this paper is to draw out a little noticed, but (I think) correct and important, consequence of David Lewis's theory of how the values of contextual parameters are determined. According to Lewis (1979), these values are often determined at least in part by accommodation; to a first approximation, the idea is that contextual parameters tend to take on the values they need to have in order for our utterances to be true. The little-noticed consequence of Lewis's way of developing these ideas is that what we say is determined in part by the way the conversation unfolds after our utterance. That is, Lewisian accommodation entails a non-standard form of externalism, according to which what we say is determined not only by factors internal to us at the time of our utterance, nor even by truths about our physical or social environment at the time of utterance or by our history, but also by truths about our future—truths about times after the time of our utterance. Seeing this consequence clearly lets us refine and improve upon Lewis's account of when accommodation can occur.

Keywords: Accommodation, context sensitivity, metasemantics, temporal externalism, David Lewis.

The purpose of this paper is to draw out a little noticed, but (I think) correct and important, consequence of David Lewis's theory of how the values of contextual parameters are determined. According to Lewis (1979), these values are often determined at least in part by *accommodation*; to a first approximation, the idea is that contextual parameters tend to take on the values they need to have in order for our utterances to be true. The little-noticed consequence of Lewis's way of developing these ideas is that what we say is determined in part by the way the conversation unfolds after our utterance.[1] That is, Lewisian accommo-

[1] I say "little noticed" rather than "unnoticed" because Mark Richard points out, in a discussion of Lewis, that "our conversational behavior presupposes that what transpires in a conversation at a time t may effect the interpretation of predicates used in contributions to the conversation completed (long) before t"' (1995: 565)—which is very close to the view I will go on to discuss. But Richard adds an important

dation entails a non-standard form of externalism, according to which what we say is determined not only by factors internal to us at the time of our utterance, nor even by truths about our physical or social environment at the time of utterance or by our history, but also by truths about our future—truths about times after the time of our utterance. Seeing this consequence clearly lets us refine and improve upon Lewis's account of when accommodation can occur.

Before I begin, let me lay out a few assumptions to ease the discussion to follow. I take a *context* to be an ordered sequence, with the elements of the sequence corresponding to specific context sensitive expressions; for example, the sequence might consist of an element corresponding to "I", an element corresponding to "you", an element corresponding to "that", an element corresponding to "tall", and so on.[2] In some cases, these elements may be the extension of the corresponding expression (e.g., the element corresponding to "I" may be an individual, the speaker), while in other cases the semantic values of the expressions may allude to these elements in some other way (e.g., we will assume that the element corresponding to "tall" is not the extension, but a degree of height—the standard that something must meet to count as "tall" in the context).

I am also going to assume that the semantic values of sentences are functions from contexts to propositions, and that these propositions serve as the content of speech acts such as assertion.[3] (So, on the view I am taking for granted, semantic values are something much like Kaplan's characters.) The idea that semantic values relate so straightforwardly to contents is controversial (Ninan 2010, Rabern 2012, Rabern and Ball forthcoming), and I am adopting it only for the sake of simplicity; nothing substantive about what I have to say would change if we adopted a different idea of what semantic values are and how they relate to content.

Since semantic facts are not brute, the values of this elements of the context will be determined by some facts about the speaker and her audience, and their environment broadly construed. Exactly which facts matter is a difficult question; this paper aims to make the case that facts about the future matter, but leaves the question of which other facts matter open. Kaplan (1989: 573–4) famously distinguishes between descriptive semantics (which aims to say what expressions mean) and metasemantics (which aims to explain why expressions have the meanings they do), and I take the question of how the elements of the context are determined to be metasemantic (perhaps in a somewhat extended sense).

complication, which (I will claim) is both unnecessary and problematic. I discuss this complication in section 2, below.

[2] For discussion of this sort of view of context, see Lewis (1970: 62–5), Braun (1996: 161), and Ball (2017: 108–9).

[3] In this respect I am being untrue to Lewis's own views; see his 1980.

1. *Lewis on Accommodation*

David Lewis (1979) defended a metasemantics on which a range of contextual factors relevant to determining the truth value of assertions—what he called the conversational score, which would include the elements of what we are calling the context—tends to shift (as Lewis says, "ceteris paribus and within limits") so as to make assertions true. Lewis calls this metasemantic mechanism *accommodation*. Lewis motivates accommodation by appeal to a range of examples; we will focus on a subset of his cases, those involving gradable adjectives like "flat" and "hexagonal". These adjectives are context-sensitive; what counts as "flat" in one situation (say, one in which we are building a road) will not count as "flat" in another (say, one in which we are sanding a tabletop). But what sets the standard? What determines how flat something has to be to count as "flat" in a given situation? Lewis's view is an attempt to give a partial answer to these questions.

To a first approximation, Lewis's idea is that if I say "France is hexagonal", that tends to make it the case that "hexagonal" as I use it is correctly applied to France (i.e., the parameter of the context associated with "hexagonal"' (call it $c_{\text{hexagonal}}$) is such that France is more hexagonal than $c_{\text{hexagonal}}$), and likewise, if I say "Hamburg is flat", that tends to make it the case that "flat" as I use it correctly applies to Hamburg.[4] He generalises these examples into the following scheme:

> If at time *t* something is said that requires component s_n, of conversational score to have a value in the range *r* if what is said is to be true, or otherwise acceptable; and if s_n, does not have a value in the range *r* just before *t*; and if such-and-such further conditions hold; then at *t* the score-component s_n, takes some value in the range *r*. (Lewis 1979: 347)

Before we proceed, we should clarify Lewis's aim in this passage. Locutions like "what is said" are often used in the literature to talk about *content*—what is asserted by an utterance. If we read "something is said" and "what is said" in the quoted passage in this way, then Lewis's idea might be paraphrased as follows: suppose an utterance expresses a certain proposition. This proposition has particular truth conditions; and it may turn out to be true just in case the conversational score is a certain way. On this understanding of what is going on, a proposition is expressed *prior to, and independently of, accommodation*, and accommodation makes it the case that that proposition is true; or in other words, first a determinate proposition is asserted and then accommodation happens.

[4] Lewis suggests that what is at issue in these examples is a "standard of precision". I am updating Lewis's treatment to be more in line with contemporary views of gradable adjectives such as Kennedy and McNally 2005. In any case, it seems clear both that there is not a single standard of precision relevant to all gradable adjectives in a context, and also that "precision" is not the right way to describe the standards relevant to many gradable adjectives. (There is no such thing as being precisely tall or precisely beautiful.)

This can't be what Lewis intended. The idea isn't that a particular content is expressed, and then the conversational score shifts so as to make that content true. (For example, suppose that contents are the sort of thing that is true or false at a world. On most views, when we are evaluating whether an assertion is true, we evaluate its content at the world in which it is made; no further element of the conversational score is relevant to this evaluation, and only in an unusual situation would we evaluate it at some other world so as to understand it as true. Of course, if I say something about the conversation—for example, that I am the speaker, or that we have adopted a strict standard for what will count as hexagonal—then there is a sense in which whether the content I assert is true depends on the conversational score. But this is a rather unusual case, and anyway it is not very plausible to think that in general the conversational score will shift to make my assertion true in this kind of case.) Rather, a better gloss on Lewis's idea is that content—what proposition is asserted—depends on the conversational score. For example, when I say, "You are a child", whether I express a proposition that is true just in case Ansel is a child or a proposition that is true just in case Magnus is a child depends on whether the element of the context associated with "you" is Ansel or Magnus.

In cases of accommodation, then, the conversational score shifts so as to make it the case that a particular, true content is expressed. For example, suppose that France is more hexagonal than c_{low}, but less hexagonal than c_{high}. Then the idea is that when I say "France is hexagonal", accommodation can make it the case that I express the proposition that France is more hexagonal than c_{low}, rather than the proposition that France is more hexagonal than c_{high}. So I take it that the schema should be read along the following lines:

> If at time *t* an assertion is made that requires component s_n, of conversational score to have a value in the range *r* if it is to be the case that a true (or otherwise acceptable) proposition is asserted; and if s_n, does not have a value in the range *r* just before *t*; and if such-and-such further conditions hold; then at *t* the score-component s_n, takes some value in the range *r*.

Accommodation doesn't always work; it isn't as though I can always speak truly by saying "France is hexagonal", no matter what. The described mechanism only operates in certain circumstances—if "such-and-such further conditions" obtain. Among the "such-and-such further conditions" are that the assertion must not be contested in the conversation; as Lewis says, "at least, that is what happens if your conversational partners tacitly acquiesce" (1979: 339). If you say "France is hexagonal" and I reply, "Yes, and Italy is boot-shaped", then the parameters of the list context relevant to both of our assertions tend to adjust in such a way that our assertions come out true; but if I reply, "No, you're wrong, its borders are actually quite irregular—just look at how Brittany sticks out", then the parameters of the context will not so adjust. For now, let's take "such-and-such further conditions" to pick out the following:

> *Such-and-such further conditions (SSFC1)* "your conversational partners tacitly acquiesce"—i.e., no one objects.

We will go on to refine SSFC1 in the next section. Before we do that, it is worth observing that even on this plain version of Lewis's view, the "such-and-such further conditions" introduce an element of backwards determination: the parameters of the list context relevant to your utterance at *t* depend in part on my reaction to your utterance after *t*. Whether you say (truthfully) that France is hexagonal-by-low-standards, or (falsely) that France is hexagonal-by-high-standards, is not determined just by you (e.g., by your intentions, beliefs, or other attitudes, or by your dispositions); it is determined by what happens after your utterance, by whether I go along with you or object.

2. Extending and Improving Lewis's Account

We should not expect an exhaustive specification of the conditions under which accommodation will take place. Even a fully developed principle along the lines Lewis sketches will only be true *ceteris paribus*; metasemantics is complicated, and we should expect that there may be exceptional cases where factors outside the scope of any given model intervene. (Who knows what will happen to the conversational score when the Martian mind-control rays strike, or the LSD kicks in?) So we should not expect to be able to draw out the further conditions in full detail.

Despite this, it is clear that we can do better than Lewis's suggestion; the matter is not as simple as (SSFC1) suggests, because it is not settled by an interlocutor's first reaction. To see this, consider the difference between the continuation of Castorp and Settembrini's disagreement in (1) and (2):

(1) *Castorp*: Hamburg is flat.
Settembrini: It is not; it has many small hills!
Castorp: Ah, I see your point. I thought that Hamburg was flat, but I was wrong.

(2) *Castorp*: Hamburg is flat.
Settembrini: It is not; it has many small hills!
Castorp: Look, of course it has some small hills. But that doesn't really matter—there are lots of reasons to think it is flat. Bicycling is easy there, etc.
Settembrini: Aha, point taken! I was mistaken: Hamburg is flat after all.

In (1), Castorp accepts Settembrini's correction. In this kind of case, I submit, it is very natural to see Castorp's initial assertion as incorrect and Settembrini's response as correct; after all, this is the considered judgment of all the parties to the dispute. In (2), on the other hand, Castorp rejects Settembrini's correction, continues to defend his initial assertion, and it is Settembrini who concedes. In this kind of case, it is very natural to see Castorp's initial assertion as correct and Settembrini's response as incorrect; again, this is what Castorp and Settembrini themselves come to judge.

Now the judgment we have just given about (1) fits well with (SSFC1). (Castorp's utterance is not accommodated—the context does not adjust so as to make him express a truth—and this fact would be explained given (SSFC1) by the fact that Settembrini objects). But the judgment we have given about (2) does not. In (2), Settembrini objects and Castorp's assertion is ultimately accommodated nonetheless—the context does adjust so as to make Castorp express a truth, despite Settembrini's objection. So whether an assertion plays a list-fixing role is determined not only by interlocutors' first responses, but by their considered judgment—by the resolution of the debate:

> *Such-and-such further conditions 2 (SSFC2)* Your conversational partners acquiesce—tacitly or explicitly, immediately or after discussion (i.e., the considered judgment of all parties to the conversation is that you were right).

Integrating (SSFC2) into an explicit account will yield something like the following:

> *The Extended Lewisian Model* If at time *t* an assertion is made that requires component s_n of conversational score to have a value in the range *r* if it is to be the case that a true (or otherwise acceptable) content is asserted; and if s_n does not have a value in the range *r* just before *t*; then: (i) if the considered judgment of the parties to the conversation is that the assertion is true; then at *t* s_n takes some value in the range *r*; but (ii) if the considered judgment of the parties tothe conversation is that the assertion was not true then then at *t* s_n takes some value outside the range *r*.

These considerations also help us see what is wrong with the suggestion (made by Mark Richard) that in cases of accommodation, we need to look at two distinct contexts: "there is every reason to say that in the sort of case we are considering, the utterance occurs in at least two contexts. For it occurs within the context established by [the speaker's] utterance at the time he makes it (we might call this the utterance's local context), and it occurs within the global context determined by the conversation taken as a whole" (1995: 566). I would argue on the contrary that the "local context" has no role substantial role to play in the story. Perhaps the clearest way to see this is by considering the metasemantics of the local context. Exactly what fixes the values of the elements of the local context? One natural proposal would be the speaker's intentions; it is unclear what other options there might be. If that is correct, then relative to the local context, Castorp asserts a truth—he is under no illusions about the topography of Hamburg, and intends to use "flat" in such a way that Hamburg counts as "flat". Settembrini is in a position to know this; so this proposition cannot be what his objection is addressing when he says, "You're wrong". (It is not as though he accepts Castorp's utterance as true and decides to object anyway; no, he thinks that Castorp is wrong, speaking falsely, and is going to try to show it.) But this leaves no work for the local context to do: it is not what the audience understands, not what is addressed even by the first response. I therefore maintain that Richard's multiplication of contexts does no work beyond that which is done by the Extended Lewisian Model, and that it should be rejected.

3. *Justification of the Extended Lewisian Model*

The Extended Lewisian Model makes good sense of the contrast between examples like (1) and examples like (2). That is interesting, but may seem a small benefit given that the view appeals to a mechanism that some may feel is extremely counterintuitive. Does the idea have anything else to recommend it?

A number of theorists have claimed that in at least some cases of dispute such as (1) and (2), at least part of what is at issue is how we should talk (see e.g. Plunkett and Sundell 2013). These theorists point out that we may in some sense agree on the facts about the topography of Hamburg—we may have the topographical map before us—and may still enter into disputes like (1) and (2). In this case, it looks like we cannot be disputing about a matter of fact. Plausibly, part of what Castorp is trying to do is to get Settembrini to use "flat" in a particular way; and likewise, part of what Settembrini is trying to do is to get Castorp to use "flat" in a particular way.

This observation is clearly compatible with the Extended Lewisian Model: if Settembrini can convince Castorp, this will play a role in making it the case that Castorp used "flat" with a particular meaning, and it seems safe to assume that this in turn will play a role in shaping his future uses (and similarly if Castorp can convince Settembrini). But there is a further datum to be made sense of: the parties to the dispute give arguments in the attempt to convince each other, and these arguments often do not bear in a straightforward way on the use of words. For example, consider Settembrini's contention that Hamburg is not flat because it has small hills, or Castorp's contention that Hamburg is flat because bicycling is easy there. These are sensible contributions to the conversation, contributions that might make us adopt particular views about the topography of Hamburg. But, except in some special cases (e.g., where are undertaking a bicycling holiday and have implicitly agreed that all and only places suitable for bicycling are to be called "flat"), they do not seem like good reasons to use the word "flat" in a particular way. There must be more to the story.

The most straightforward way to make sense of conversations like (1) and (2) is that the parties to these conversations are giving arguments, trying to provide (at least pro tanto) reasons to believe some conclusion; and that at least in many cases these are *good* arguments. Now, of course it isn't that we want every argument anyone ever gives to be a good argument. We sometimes make mistakes; in many cases, these may go by undetected, but in others we will look back on our own arguments and find them wanting—for example, as we imagine Settembrini doing in (2). But in many cases, we look back on our own arguments and find no fault with them. Ideally, we should want a view that vindicates our considered judgments about our arguments.

I claim that the Extended Lewisian Model does exactly that. It makes our arguments good in the following sense: to the extent that we are rational, when we look back on a dispute that has resolved, the arguments that we take to be good will in fact be good, and the arguments we take to be bad will in fact be bad. To get a sense of why this should be so, let's look more closely at the exchange that begins (1):

(3) *Castorp*: Hamburg is flat.
Settembrini: It is not; it has many small hills!

At the beginning of the conversation, Castorp intends to use "flat" in such a way that Hamburg counts as "flat", the fact that a city has small hills is no reason (or at most a very weak reason) to think that it is not "flat", and the fact that bicycling in a city is easy is a good reason to think that it is "flat". Settembrini, by contrast, intends to use "flat" in such a way that the fact that a city has small hills is a good reason to think that it is not "flat", and (hence) that Hamburg is not "flat". Of course, given the Extended Lewisian Model, these intentions are not decisive; so we do not have enough information to say whether Settembrini's argument is a good one. If the argument continues as in (1):

(4) *Castorp*: Ah, I see your point. I thought that Hamburg was flat, but I was wrong.

Then Castorp and Settembrini will look back on Settembrini's argument as a good one; and given what "flat" means (and meant, even in Castorp's initial utterance), the argument will in fact be a good one. If, on the other hand, the argument continues as in (2):

(5) *Castorp*: Look, of course it has some small hills. But that doesn't really matter—there are lots of reasons to think it is flat. Bicycling is easy there.
Settembrini: Aha, point taken! I was mistaken: Hamburg is flat after all.

Then both parties will look back on Settembrini's argument as a bad one; and given what "flat" means (and meant all along), it will in fact be a bad one. (And similarly both parties will look back on Castorp's argument to the effect that Hamburg is flat because cycling is easy there as a good one, and so it will be.) So Backwards-Looking Meta-Contextualism vindicates exactly those arguments that the disputants take to be vindicated at the end of the dispute.

4. *Conclusion*

The extended Lewisian meta-semantics presented here thus does a good job of making sense of the way we argue and evaluate our own arguments, while also vindicating the idea that many debates turn on questions of meaning. No doubt it raises further issues; but exploring these is a task for further work.[5][6]

References

Ball, D. 2017. “What are we doing when we theorise about context sensitivity?” In J. J. Ichikawa (ed.), *The Routledge Handbook of Epistemic Contextualism*. London: Routledge.

Ball, D. Forthcoming a. “Revisionary Analysis without Meaning Change, Or, Could women be analytically oppressed?” In A. Burgess, H. Cappelen, and D. Plunkett (eds.). *Conceptual Ethics and Conceptual Engineering*. Oxford: Oxford University Press.

Ball, D. Forthcoming b. “Relativism, metasemantics, and the future.” *Inquiry*. Forthcoming in a special issue edited by Henry Jackman.

Braun, D. 1996. “Demonstratives and their linguistic meanings.” *Nous* 30: 145–173.

Jackman, H. 1999. “We live forwards but understand backwards: Linguistic practices and future behavior.” *Pacific Philosophical Quarterly* 80: 157–177.

Jackman, H. 2005. “Temporal externalism, deference, and our ordinary linguistic practice.” *Pacific Philosophical Quarterly* 86: 365–380.

Kaplan, D. 1989. “Afterthoughts.” In J. Almog, J. Perry, and H. Wettstein (eds.). *Themes From Kaplan*. New York: Oxford University Press: 565–614.

Kennedy, C. and McNally, L. 2005. “Scale structure, degree modification, and the semantics of gradable predicates.” *Language* 81: 345–381.

Lewis, D. 1970. “General semantics.” *Synthese* 22: 18–67.

Lewis, D. 1979. “Scorekeeping in a language game.” In *Philosophical Papers Volume I*. New York: Oxford University Press: 233–249.

Lewis, D. 1980. “Index, context, and content.” In *Papers in Philosophical Logic*. New York: Cambridge University Press: 21–44.

Ninan, D. 2010. “Semantics and the objects of assertion.” *Linguistics and Philosophy* 33: 355–380.

Plunkett, D. and Sundell, T. 2013. “Disagreement and the semantics of normative and evaluative terms.” *Philosophers’ Imprint* 13 (23).

Rabern, B. 2012. “Against the identification of assertoric content with compositional value.” *Synthese* 189: 75–96.

Rabern, B. and Ball, D. Forthcoming. “Monsters and the theoretical role of context.” *Philosophy and Phenomenological Research*.

Richard, M. 1995. “Defective contexts, accommodation, and normalization.” *Canadian Journal of Philosophy* 25: 551–570.

[5] See Jackman 1999, 2005 and Ball (forthcoming a, forthcoming b) for a start to this further work.

[6] My thanks to the audience at the Dubrovnik Philosophy of Linguistics workshop in September 2017, to audiences in Oslo and Buenos Aires, to Torfinn Huvenes, and to Mark Richard.

Croatian Journal of Philosophy
Vol. XVIII, No. 54, 2018

Predicates of Personal Taste: Relativism, Contextualism or Pluralism?

NENAD MIŠČEVIĆ
University of Maribor, Maribor, Slovenia
Central European University, Budapest, Hungary

The paper addresses issues of predicates of taste, both gustatory and aesthetic in dialogue with Michael Glanzberg. The first part briefly discusses his view of anaphora in the determination of the semantics of such predicates, and attempts a friendly generalization of his strategy. The second part discusses his contextualism about statements of taste, of the form A is Φ, and then proposes a pluralist alternative. The literature normally confronts contextualism and relativism here, but the pluralist proposal introduces further options. First, it distinguishes first-level and second-level, more theoretical, approaches. At the first level it introduces the naïve view option, the naive non-dogmatist experiencer who simply claims that A is Φ and that's it. On meta-level such an experiencer is simply agnostic about further matters. Then, there is the first-level dogmatist stance, characteristic for people who do sincerely debate the issues, who naively believe they are objectively right. The third option is the tolerant, liberal one: "A is Φ; for me, I mean. How do you find it?" On the meta-level, dogmatic disagreement goes well with value-absolutism, entailing that one of the parties is simply wrong, and with relativism. If one is not dogmatist about taste predicates, one should accept that dogmatist is simply wrong; no faultlessness is present. The liberal stance goes well with contextualism. If one is liberal there is no deep disagreement. So, the idea of faultless disagreement is a myth. But the proposal notes that language is open to all possibilities, there is no single option that is obligatory for all speakers.

Keywords: Predicates of taste, relativism, contextualism, pluralism.

1. *Introduction*

"Chocolate is tasty", "Rollercoasters are fun"; such seemingly simple sentences and judgments have become a widely discussed topic in philosophy of language, of art and elsewhere. These will be our topic in this paper; I hasten to add judgment of aesthetic or artistic taste, like "Matisse is better than Picasso," (see Young 2017: 108). Some authors talk about "sentences expressing subjective judgment" (Lasersohn 2017: 1), and then list judgment of taste (…is fun, …is tasty…), and other judgments expressing evaluation (…is good, …is beautiful). We shall focus on judgments of taste, both gustatory and slightly more general, let us call it "hedonic" (…is fun), and then apply the idea to the aesthetic taste and to aesthetic judgments.[1]

Glanzberg's theoretical ambition is to offer a unitary truth-theoretic semantics for such judgments. He opts for one approach, the contextualist one, rejecting relativism and other alternatives. I must note at the very beginning my debt to Glanzberg. I shall discuss his brilliantly defended proposal, and then propose an alternative, indeed a pluralist one, claiming that the sentences in question can, and often do, express different judgments in the mouths of different person. A child might claim that "chocolate is tasty" and that "rollercoasters are fun", period, finding others who disagree simply not worth of attention. But n the course of time the child might learn that others she cares about have opposite opinions, and realize that, well, chocolate is tasty-for-him-and his likes. I shall argue that she is thereby learning both about the world and the language.

So, here is the preview. The rest of the present section introduces the taste predicates, and also a related notion of response-dependence. Then we turn to questions inspired by Glanzberg. Section two follows Glanzberg applying the semantics-pragmatics distinction. We take over his analysis of anaphora, as the symptomatic mechanism that guides the constitution of taste-related meanings and their understanding. Then we very briefly, with apologies, attempt to widen the model to other possible uses of anaphora, as a guidance from syntax-cum-semantics to issues of reference (and truth-conditions) determining in the context. Anaphora enables us to widen the semantic foundations for such determining, against extreme pragmaticist, who would make it completely pragmatic. Section three turns to Glanzberg's contextualism about predicates of personal taste. I find it to be a correct description of one possible attitude connected with taste, but I think there is no reason to be dogmatic about there being a single correct attitude. So, in the next to last section I turn to the pluralist alternative, trying to integrate the main options from the literature, and offer additional characterization, ending thus with six characterization, that can be mutually combined to yield more precise description of how individuals use and understand predicates of personal taste. I also briefly indicate how the theory might be extended to issues of taste that go beyond gustatory and hedonic taste, for instance in the direction of artistic-aesthetic taste. The whole spectrum of options is again summarized in the Conclusion.

[1] The present paper originated as a comment on Michael Glanzberg, for Philosophy of Linguistics and Language conference at Interuniversity Center Dubrovnik, Sept. 2017. Thanks go to Michael, and to Dunja Jutronić.

Let us start with elementary mattes. Following Glanzberg we shall concentrate on predicates having to with gustatory taste (The food in restaurant *Orhan* in Dubrovnik is tasty.), and with a wider area that might be called "hedonic taste", exemplified by adjectives such as "fun" (Reading Glanzberg is fun). Also following Glanzberg, we shall talk of parameters, semantic and pragmatic, stressing the parameter of experiencer (the reader who finds reading fun) or judge (a third person who judges that reading Glanzberg is fun). Glanzberg is more into stressing the role of experiencer, other, for instance Lasersohn (2017) is more sympathetic to judges. Besides relying on two Glanzberg's papers (2007, 2016), I shall occasionally refer to Lasersohn, above all to his recent (2017) book which I find quite congenial.

Let me briefly mention two issues that will accompany us throughout the paper. First, there is a respectable tradition (Wright 2008, Lasersohn 2017, Kölbel 2011...) that sees the disagreement in matters of taste as faultless disagreement. Glanzberg is against it (2006:16), and I tend to agree with him. I think that for disagreement to be genuine, the participants have to be dogmatic about their taste(s). If they are naive non-dogmatists the disagreement does not arise. Same if they are liberal. But if they are serious dogmatists, they are both wrong! No faultless disagreement, or so I shall argue.

The other issue is the relation to response-dependence, also noted by Glanzberg. Here is a reminder of Hume, who clearly connected taste with response-dependence:

> 'Tis a common observation, that the mind has a great propensity to spread itself on external objects … (Hume 1978: I.iii.XIV).
>
> Taste has a productive faculty, and gilding and staining all natural objects with the colours, borrowed from internal sentiment, raises in a manner a new creation … (Hume 1983: 88.)

There are many areas in which basic properties might be response-dependent: color, aesthetic objects, emotional qualities (sadness of a situation), meaningfulness of a situation or even of a life as a whole, and then morals. My own conjecture is that most of the manifest properties in human world are response-dependent. Here, I shall just note the connection with predicates of personal taste, and leave the further investigation for future.

2. *Glanzberg: semantics, pragmatics and guidance*

2.1 *Glanzberg's proposal*

First, a brief methodological question. What constitutes the meaning of a sentence or a statement and how does one find out its semantics? If Nenad says "Reading Glanzberg's paper is fun", what determines the full meaning of the statement, and how do we recognize it?[2] Glan-

[2] Devitt (2013) rightly warns against confusing the two questions, the constitutive and the epistemological one.

zberg discusses the issues as metasemantic ones (and reminds us that metasemantics studies how semantic values, including context-dependent ones, get fixed (2016: 2). Here is his summary:

> II.3 Metasemantics of *E*
>
> - In section I.2 I noted that implicit thematic arguments have what I call a direct metasemantics. Recall, metasemantics describes how the semantic value of an expression gets set (including how context can help to set it). Direct metasemantics is on the model of a demonstrative, where a referential intention of the speaker in effect directly sets the value. Thematic positions tend to go with referential intentions on the part of speakers. These are especially important for context dependent arguments, where the referential intention does a substantial share of the work in setting the value. We thus see that thematic arguments, even implicit ones, have direct metasemantics. We shall see that this holds for *E* as well, though with a small but I think interesting qualification. (2016: 32)

Glanzberg rightly sees syntax as a guide for semantics. If a trait is recognized by syntax, then it is semantic, not pragmatic. For instance, writing about focus, he notes that

> [i]t provides cases where what appears to be surface syntax is not a good guide to underlying linguistic form. This lesson has been learned before, but focus shows that what is on the surface but appears to be merely pragmatic can turn out to indicate underlying syntactic structure. Association with focus shows that this structure can be semantically significant. The first moral of focus is that the appearance of being merely pragmatic can drastically deceive. (2005: 106)

An interesting application of this idea is his stress on anaphora as the indicator of meaning (2016: 30). Start with sentences

> "(34) "Bill, Max, and I were eating duck tongue in the market. It was tasty".

Tasty to whom? People find the reading "tasty to us" as fine. So, one should take "It was tasty" as pointing to the experiencer(s), reference to whom is hidden in the preceding sentence "Bill, Max, and I were eating duck tongue in the market." And he notes that it is a clear case of anaphora.[3] And he rightly comments:

> II.2.2 Anaphora So far, I have noted that the experiencer of a predicate of personal taste is somehow marked for point of view. This is to give a name to a mysterious phenomenon, but it at least points out some substantial semantic restriction on the value of *E*, if not more. I have labeled point of view a semantic phenomenon, as it appears to be a standing feature of the meaning of an argument position. But it also relates to what kinds of values it can have in context, and so affects pragmatics. (2016: 29).

[3] And he rightly finds the anaphora-tied explanation in the cases that are less clear to hearers. Here is the quote:

(33) Three people were eating duck tongue in the market.

a) Susan looked (= looked at them).

b) OK/? It was tasty (= tasty to them).

Most of my informants found (33b) acceptable, but many found it somewhat degraded, and more found it clearly degraded in comparison with (33a)" (2016: 30). Since there is no clear anaphora, hearers find it degraded, and have trouble recognizing the experiencer.

I think it is a particular instance of the phenomenon of anaphora and guided saturation that is practically omnipresent in normal conversation.

Pragmaticists, like Francois Recanati (see his 2010 book), have argued that most typical sentences used in everyday speech are incomplete and should be saturated by free pragmatic interpretation, in order to yield truth-evaluable contents. For instance, "I've had breakfast" is truth-conditionally indeterminate (when?), and "It's raining." As well as where is it raining. Or take "Everybody went to Paris", who counts as everybody? But such sentences would most often be said in situations of ongoing communication where a question has been asked, or a pointing has been made, and the like. Consider

> I've had breakfast.

When are such sentences normally produced? Often as an answer to the question "Did you have breakfast today?", or "You look hungry, did you have breakfast?" If we incorporate the question, we end up having the truth-conditional content. "You look hungry, did you have breakfast?" is a usual question, and everyday knowledge about periodicity of having breakfast indicates that what is meant is "today": we again have the truth-conditional content.

Finally, and most ironically, the weather reports. Actual weather reports on TV give you a map with rather precise contours! But with ordinary statement of

> It's raining.

there is a rich area of possibilities. It could be a comment out of blue. Then, *and only* then is it indeterminate the way our authors see it. Normally, it could be and often is an answer to a question:

> "What is the weather like in Budapest?",

and then we have an anaphora. Or,

> "Take a look. Is it snowing?" ...is it sunny?"

It is quasi-anaphoric and suggests the area immediately surrounding "you".[4]

2.2 *A general proposal: anaphora and guided saturation*

Indeed, the clearest case of semantic determination from the context of conversation is the case of anaphora. I would like to generalize the mor-

[4] Glanzberg also suggests an interesting, and for him highly relevant consequences to be derived from the role of anaphora:

> The evidence from anaphora also offers a consideration in favor of treating the point of view restriction as a simply a restriction on the values *E* can take—like the content of the indexical I—rather than as writing *from the point of view of* into the content of the experiencer argument position. (2016: 31)

We cannot discuss it in any detail here.

als of our examples, independently of the case of predicates of taste.[5] Let me use material from Stainton; he is more of a pragmaticist, so I cannot be accused of using biased evidence from linguistics. Here is how he introduces the case of elliptic speech exemplifying anaphora.

> Imagine Steve being asked the question
> 'What language does Mary write in?'
> and he says: "In Latin". Steve obviously believes that Mary writes in Latin, and in addition, he thinks that having said "In Latin" he has suggested that Mary writes in Latin. The phenomenon is known as anaphora, and speakers normally have no problem with it. If one asked Steve what he informed his interlocutor about, he would produce the judgment "that Mary writes in Latin." (Stainton 2006: 33)

Here we have the example of anaphora, where the work is done by the syntax, and the resulting intuition that the meaning (in the example "Mary is writing in Latin" is semantic. In looser situations, where there is no pronounced antecedent, the listener and the judge are guided by the canon of the strict case. Let me use Stainton's example (from Stainton 2006: 34). Rob points to a boat and says "Pretty fast". Stainton contrast this with the stricter situation, in which there is a clear syntactic antecedent (He presumably has in mind cases like the following short conversation: "How do you find the speed of the boat? Pretty fast").

But note that *the looser situation is analogous to the stricter situation* What did Rob say? That the boat is pretty fast, intuition replies. And it is *almost* correct. It seems that the listener's and judge's intuitions proceed *by analogy*.

We can then use anaphora as the model for partial determination or guidance: it assumes that anaphora is semantic, shows that anaphor guides the hearer in determining the truth-conditional content, argues that most problematic cases are anaphor-like (and the rest can be dealt with). Let me call my proposal "guidance view". The main steps are easy to grasp:

> First, assume that anaphora is semantic,
> Second, show that anaphora determines the truth-conditional content guides the hearer in recognizing the determination and the content,
> Third, argue that most problematic cases are anaphora-like (and the rest can be dealt with) and that in quasi-anaphoric cases the hearer proceeds in an analogical fashion. If this holds, it follows
> Fourth, that guidance view is very close to being the right one.

By guidance I here mean objective guidance, or quasi-determination, not mere epistemic help. I admit that the construction of content is literally and *stricto sensu* pragmatic, but it is strongly determined-guided by semantic elements.

Let me summarize. I am taking anaphora as a model, very much in line with Glanzberg. I also assume the following distribution of situations. First, complete out-of-the-blue utterances are extremely rare in

[5] With thanks to Glanzberg, and also to Michael Devitt and Robert Stainton, with whom I have discussed it.

normal situations, and when they appear, they are typically uninterpretable. Second, normal anaphora is quite frequent: elliptic sentences very, very often appear in reaction to the preceding discourse, which completely determines the slot-filling. In between these two extremes, we find quasi-anaphoric situations, in which there is a verbal antecedent, but it does not clearly determine the slot-filling in a linguistically-semantically unproblematic way. Finally, there are situations in which verbal antecedents are replaced by other events in interaction: the common direction of the gaze, pointing or almost pointing gestures, and the like.

A pragmaticist, like Stainton, might attack it from the opposite direction: the "guidance" is just a pragmatic phenomenon that has little or nothing to do with semantics. But consider the analogy with indexicals. Their content was first seen as pragmatic, but soon, already with Montague, theoreticians recognized the strong and systematic determination of content, and started counting saturation for main indexicals (I, now) as being semantic or almost. I propose we do the same with quasi-anaphora: the proximity to the pure syntactico-semantic determination (i.e. proximity to anaphora *stricto sensu*) suggests an analogous semantic treatment.

3. *Glanzberg's contextualism about predicates of personal taste*

Here is Glanzberg's official contextualist proposal. He argues that

> [P]redicates like tasty and fun are context-dependent is not all that controversial (...). At least, these expressions show some of the same context dependence that other predicates built from the positive forms of gradable adjectives do:
> context helps to somehow set the standard for how tasty or fun something has to be to count. Just as someone can count as tall relative to one context, where jockeys are under discussion, and not tall in another context, where basketball players are under discussion; so too a cheeseburger might count as tasty, relative to a context where bad bar food is under discussion, and not tasty, relative to a context where the best foods in California are under discussion. (2007: 9)

And he notes:

> The semantics I have just sketched is a 'contextualist' one, attributing the interesting properties of predicates of personal taste to context dependence. This stands in contrast to recent relativist analyses of these sorts of predicates… (2016: 17)

> However the claims I shall defend here are mostly orthogonal to the fundamental points of contention between relativist and contextualist accounts. (2016: 18)

However, in his (2007) paper he has argued against relativism, presenting and defending a contextualist semantics; here, I shall take both papers into account.

> I shall argue that predicates of personal taste, like fun and tasty, contain two hidden contextual parameters. One is the familiar standard parameter, which (...) is a functional parameter. The other is the experiencer parameter I have claimed is present in these predicates (Glanzberg 2007: 15)

And here is his introduction of the experiencer parameter *E*

> Adjectival predicates of personal taste, like our paradigmatic tasty and fun, are gradable adjectives, and so have a standard parameter. But what makes them personal taste predicate, I have claimed (Glanzberg 2007), is that they also have an experiencer parameter, which I label *E*. That in turn acts the sort of scales they use, in such a way as to interpret them as being about the personal tastes of the experiencers. I have thus argued that the semantics of these predicates looks something like:
>
> (15) a. $[[\text{tasty}]]^{c}$ = degree-gustatory-quality-experienced-by-*E* = $\lambda x.\text{tasty}_E(x)$
> b. $[[\text{fun}]]^{c}$ = degree-enjoyment-experienced-by-*E* = $\lambda x.\text{fun}_E(x)$ (Glanzberg 2016: 17)

I shall later argue that the picture corresponds to the the liberal attitude, call it *E*-liberal one.[6] And he continues:

> This stands in contrast to recent relativist analyses of these sorts of predicates, starting notably with Lasersohn (2005). I of course, think the contextualist view is correct, and the version of it i have defended relies on these contextual parameters. .. indeed, the claims I shall defend here are mostly orthogonal to the fundamental points of contention between relativist and contextualist accounts. (2016: 17)

It is a pity that he does not discuss the contention in the same (2016) paper, so we had to rely on the earlier, (2007) one. He also offers a lot of syntactic evidence; we have to skip it, unfortuntely. So, we stay with semantics and pragmatics of *E*. Glanzberg notes that the *E* is clearly a thematic argument, assigned an experiencer thematic role by

[6] Here is more on standard setting parameter:

we are assuming that there is some kind of hidden contextual parameter in gradable predicates that sets a standard, e.g. for tall, a standard for how tall something has to be to count as tall. Following Kennedy (2007), we considered a couple of options for how that might work (...):

(52) a. Max is tall.
b. Tall (Max) > d_c.
c. Tall(Max) > s(tall)

In the first, we simply have a contextually provided standard value d_c, in the second, we have a contextually provided function that computes the standard for a given adjective. (2016: 42)

Combining the two parameters, for an occurrence of a predicate of personal taste in positive form, gives something like:

(16) a. Stewed duck tongue is tasty.
b. tasty_E (Stewed duck tongue) > $s(\text{tasty}_E)$

For our purposes here, the most important feature of this analysis is that it relies on both the contextual parameters ***s*** and *E*. The presence of ***s*** is widely accepted (...). The claim that we need an experiencer parameter *E*, on the other hand, is in more pressing need of defense, ...

The semantics I have just sketched is a `contextualist' one, attributing the interesting properties of predicates of personal taste to context dependence. (2016: 17) the predicate, and calls it `thematic hidden parameter' (2016: 27). Most importantly, the experiencer argument picks up its value from the context, usually include the speaker. Most importantly, the parameter is not part of the content, it is *„not written into the proposition expressed."* (2016: 28).This is the basis of Glanzberg's contextualism about taste predicates.

Let me conclude this brief, all too brief, summary of Glanzberg's views noting that he does take seriously the response dependence that is probably linked to taste and taste properties

> To say there is some such parameter for an experiencer is not to determine how it affects the interpretation of expressions like fun. Presumably, if the experiencer class is not inert, we should see some sort of response dependence in the meaning of fun—where the experiencer class fixes whose responses count. But, how the experiencer class does this, and how much response dependence we see, remain questions.
>
> There are lots of properties which have a significant degree of response- dependence, but are not fully response-dependentist. (2007: 12)

Let me note that the candidate area for response-dependence is extremely large. Start with color, say, orange. The response dependentist suggests that being orange in objective sense is being such as to cause the response of visaging phenomenal orange in normal observers under normal circumstances.[7]

We shall be briefly mentioning a taste-related property, beauty. Here again, the response dependentist claims that being beautiful in objective sense is being such as to cause the response of visaging phenomenal beauty in normal observers under normal circumstances.[8] As noted in the Introduction, there are more areas: emotional qualities (sadness of a situation), meaningfulness of a situation or even of a life as a whole, and then morals. My own conjecture is that most of the manifest properties in human world are response-dependent; if this holds, and if in many cases response-dependence has a tight connection with taste, there might be a lot of work to do along the lines briefly alluded to by Glanzberg.

[7] The standard form of response dependentist argument for this conslusion can be very briefly summarized in the following way

Full phenomenal orange is being intentionally experienced as being on the surface of the fruit. (*A transparency datum*)

Full phenomenal orange is not on the surface of the fruit. (*From science*)

Full phenomenal orange is not a property of subjective state (*From Transparency*).

Therefore (*by principles of charity and by inference to the best explanation*)

The above conclusion follows.

[8] The form of the argument is the following:

Beauty (phenomenal) is being intentionally experienced (visaged) as being a property of the picture.(*A transparency datum*)

Beauty is not a viewer-independent property of the picture. (*From science*)

Beauty is not a property of subjective state (*From Transparency*).

Therefore (*by principles of charity and by inference to the best explanation*)

The above conclusion follows.

But let us return to our specific topic, the semantics of taste predicates. Here, I completely agree with Glanzberg that his contextualist proposal offers a fine reading of many sentences involving predicates of personal taste. My question is whether this reading is the only one. For instance, I hope that my infant grandson will be able, in a year, to say "chocolate is tasty". He will thereby indicate to me that he finds chocolate tasty, but I seriously doubt that he has any reflective knowledge of parameters that might be relevant. For him, being tasty is just the property of chocolate. This does not entail that *he does not understand his language*.

For the example of an opposite situation, still within my family, let me turn to my wife and myself; I have sweet tooth, and, like my grandson, I love chocolate. My wife is not attached to sweet things; she would always prefer fresh fruit to chocolate. When we talk to each other, each of us takes these things into account; when I say "chocolate is tasty", I don't mean is should be tasty for her. And *vice versa*. Again, each of us has a good mastery of language. So, the use of the sentence is slightly different between us (the grandson at his future stage included).

So, why insist that there is just one reading of the sentence? Why not turn to a pluralist alternative?

4. *A pluralist proposal*

I doubt that there is a single correct reading of the use of taste predicates along the lines of *any one* of the proposals we looked at. I agree with Glanzberg that on a sophisticated reading (that I will call „liberal") *E* determines ***s*** in the context, but this does not *dictate* the self-understanding of the speaker. In other words, the relevant sentence (say, "*Roller-coasters are fun.*") admits of several meanings and interpretations, and can express several judgments, some more relativist and some more contextualist, some more dogmatic and others more liberal.

Consider the options again. We have three immediate options in relation to a statement of taste, of the form A is Φ (e.g. "*Roller-coasters are fun.*") and to the stance taken by the speaker-experiencer:

1. *naive non-dogmatist* experiencer who simply claims that A is Φ and that's it. On meta-level such an experiencer is simply agnostic about further matters, like weather A is Φ for other people, who is right about it, and so on.
2. a bit more reflective stance is the *dogmatist* one: If you don't agree, you just don't know about A being Φ. I think people who do sincerely debate the issues are honest dogmatists, who naively believe they are objectively right,
3. the tolerant, liberal one: „A is Φ; for me, I mean. How do *you* find it?"

Glanzberg's official claim: *E*, and ***s*** determined by *E* are *not* part of the content of speaker's claim can go with both 1 and 2. The naive non-dogmatist experiencer think she is just describing the way roller-coasters are (The way A is). The dogmatist re-interprets *E* (so to speak) as being universal in scope. Consider:

. John-the-dogmatist says: "Roller-coasters are fun."
Mary: But they are not fun.
John: You are dead wrong.

Their being fun is just seen by him *as a fact* (a value laden fact), not as something that is due to his perspective. Mary can continue, stressing the difference in judgment:

Mary: Sorry, *you* are dead wrong.

Here we have genuine disagreement, and if there is no universal norm for being funny, the disagreement is not faultless. (For Lasersohn disagreement comes with judge or opinion parameter; but is the parameter essentially different from *E*?)

So, we have a dogmatic option: if (disagreeing) speakers are dogmatic, *E* and ***s*** make no appearance in the content; the content is just that roller-coasters are fun, period. Or that they are not fun, period. And the disagreement is far from being faultless. The absence of *E* and ***s*** from the content looks good for disagreement, bad for liberal tolerant spekers.

Consider now the relativist alternative: *E* (or some "judge"-parameter) and ***s*** determined by *E*_are part of the content of speaker's claim. (against the official Glanzberg's claim). The alternative is compatible with two options.

Option one dogmatist. John takes his claim, namely the content "roller-coasters are fun" as *the truth*.

Option two, liberal. If Mary disagrees, John will respect her claim; anyway, the content of his opinion is not essentially tied to his perspective**.** The situation is parallel with the standard use of indexicals:

- John: I am hungry
- Mary: I am not hungry.

John accepts a non-absolute status for his claim. He agrees that "Roller-coasters are fun" usually means fun for the speaker or the speakers' salient group of friends or family. It usually means fun for me or fun for the whole family. But this is our liberal, tolerant option. And no disagreement with Mary. Roland Barthes gives a fine example of a universalist liberal attitude (he does not call it like this) in his retelling of Fourrier's predilections: "the society cannot rest until it has guaranteed (…) the exercise of my manias, whether bizarre or minor" (1989: 77); his example involve liking rancid couscous, linking old chickens and eating "horrid things", like for example the astronomer Lalande eating live spiders. Taste is not to be commanded (1989: 77).

In the literature, relativism is connected with disagreement, and the latter is characterized as faultless.[9] But at the same time, relativism claims the proposition affirmed in John's utterance has a truth value only relative to John's standards, when he is the assessor, and Mary's standards, when Marry is the assessor. This gives one some disagreement, but it is hard to see how it can be faultless. From a liberal contextualist standpoint both dogmatic relativists are at fault, so John's disagreement with Mary is not faultless.

In order for the speaker to be non-dogmatic, (s)he has to accept the validity of other points of view and the ***s***'s that go with them.then, in short, (s)he has to be liberal:

- *Roller coaster is fun* for me, you know. But, how do you find it?
- I find it not fun, for me, I mean, but I understand your predilection!

But then, the disagreement is lost, and the explanation of disagreement is lost, as Glanzberg also noted (2006: 16).

What about theoretical perspectives accompanying the three first-order stances? Consider it case by case. The dogmatic disagreement goes well with value-absolutism, entailing that one of the parties is simply wrong, and with relativism. Here, it seems to me that no faultlessness is admitted by the speaker; his interlocutor is at fault. If one is not dogmatist about taste predicates, one should accept that dogmatist is simply wrong; no deep disagreement is present. Such a liberal stance goes well with some versions of contextualism. If one is a liberal there is no deep disagreement, so, the idea of faultless disagreement is a myth. In this case, liberalism is wiser than dogmatism. So, I find the whole idea of faultless disagreement dubious: if the speaker is dogmatic and disagreeing there is no faultlessness, if she is liberal, non-dogmatic there is no real disagreement. Here I agree with Glanzberg who once described the idea of faultless disagreement as "absurd" (2007: 16).

But note that language is open to all possibilities. The language of taste attitudes is compatible with all three first-order stances: with naive non-dogmatism, with dogmatism and with tolerant liberalism. Particular uses of language can be classified along second-order options, as agnostic, absolutistic, relativistic and contextualist. But the whole business is linguistically correct, syntactically, semantically and pragmatically, so *there is no single correct reading of the use of taste predicates and the like.* Our naïve agnostic is linguistically in the clear. On the other hand, the absolutist does not reform language, she is into postulating objective value-properties in the world. The relativist is not making a linguistic mistake; her being right or not depends on the domain which is being judged. We are dealing not with semantics, but with matters of reality!

[9] See, for example, the sources mentioned in the Introduction: Wright 2008, Lasersohn 2017, and Kölbel 2011, as well as papers collected in García-Carpintero and Kölbel 2008.

Now, why do people debate questions of taste? Lasersohn, for example, offers two mutually contradicting answers. First, a cognitive rationale:

> Two parties will normally engage in a dispute about a matter of taste only if each of them regards the other as making an error of taste. This in no way represents a retreat from the idea that disagreements over matters of taste are faultless in our original sense, but is simply a clarification of what kind of fault was envisaged. (Lasersohn 2017: 210)

But then he also offers a pragmatic-sociological rationale:

> the point of the parties in dispute is to gain a social advantage for one's own tastes: to have them adopted more widely, or to give them priority over the tastes of others in planning and decision-making (Lasersohn 2017: 211)

and, importantly, for him this is unconnected with ascribing error to the opposite view!

> "Prevailing" in such disputes cannot mean showing that one's opponent has made some error of fact or logic. The purpose of pressing a dispute over matters of taste is to gain a social advantage for one's own tastes: to have them adopted more widely, or to give them priority over the tastes of others in planning and decision-making. (Lasersohn 2017: 211)

So, for John and Mary to engage in such dispute, it is crucial that John regards Mary as making an error of taste and vice versa. But then we are told that "/p/revailing" in such disputes cannot mean showing that one's opponent has made some error of fact or logic. The two claims simply don't fit together.

My guess is, of course different. I don't agree with the sociological rationale, and I prefer the cognitive one. I think that people who do sincerely debate the issues gustatory or hedonic taste are dogmatist (for example, *E*-relativists), or absolutists who naively and honestly believe they are objectively right. However, as Lasersohn noted—see for example chapter *10.1 Aesthetic judgment and refinement of taste*, of his (2017) book—various response-dependence linked adjectives can and do vary in the degree of dogmatism their standard use allows or requires.

Let me conclude the part on gustatory taste by a summary in form of a table.

	TASTE ATTITUDES—LANGUAGE IS COMPATIBLE WITH ALL OF THEM		
EXPERIENCER	*NAÏVE NON-DOGMATIST*	*DOGMATIST*	*LIBERAL*
1ST ORDER VIEW	Roller-coasters are fun, that's it.	If you don't agree, you just don't know what real fun is.	Roller-coasters are fun, for me, i mean. How do *you* find them?
META-THEORY	*AGNOSTICISM*	*ABSOLUTISM OR RELATIVISM*	*CONTEXTUALISM*
NO LINGUISTIC DICTATE:	Our agnostic is linguistically in the clear...	The absolutist does not reform language and the relativist is not making a *linguistic* mistake...	Finally, the contextualist is in clear...

Most importantly, language is open to all possibilities**;** there is no linguistic dictate.

What about other response-dependent areas? Here, Lasersohn is a good guide. We are lax about gustatory and hedonic taste, but less so about emotional properties: if someone finds the death of a child comic, we shall be condemning the person. Other response-dependent predicates in other areas might behave similarly.

A nice case is aesthetic-artistic taste that might be more dogmatic: professionals in the field tend to be such about their opinions: Matisse is either better than Picasso, or equal or worse, and if you have a good artistic taste you will agree with me!! They normally don't take their disagreements to be faultless, nor are they normally liberal about their judgments. Different taste areas might have different levels of objectivity taste in flavors might be completely subjective, but in other areas a more dogmatic approach might better capture the actual structure of the relevant value. As Lasersohn notes "certain perspectives may be ranked as objectively better than others" (2017: 214). He mentions that "claims about future contingent events later perspectives seem better than the earlier ones"; the same for epistemic modals in general (2017: 224). Similarly for art, some perspectives are better that others: Matisse-lovers might be right and Picasso-lovers, like the present author, might be wrong.[10]

[10] Some rare critics might be non-dogmatic, for instance, Clive Bell who wrote

Such an understanding could bring together three independently plausible ideas. The first is that beauty is response-dependent: being beautiful in objective sense is being such as to cause the response of experiencing phenomenal beauty in normal observers under normal circumstances. The second is that there is some degree of objectivity about beauty (and artistic value in general). And the third is that judgments of pictorial beauty are judgments of taste, with all the accompanying semantic options. James O. Young recently noted:

> The question of whether aesthetic judgements are simply statements about subjective preferences or whether they have some non-subjective basis is one of the most important questions of aesthetics, and, indeed, of philosophy.
> Despite the importance of the question, it has received fairly little attention in recent years. (....) A large majority of philosophers of art is opposed to subjectivism, but comparatively few contemporary aestheticians have argued against it or for a contrary position. Philosophers of language have considered aesthetic judgements, but they have tended to assume that some form of subjectivism is correct. (Young 2017: 1)

It is enough to stick to more dogmatic reading of judgments of aesthetic-artistic taste: Either Matisse-lover is right or Picasso-lover is right; some perspectives are better than others so, no relativism follows. Some response-dependent properties allow for objective standards, so, let us hope the aesthetic-artistic and moral properties are such. But this is a topic for another occasion

5. *Conclusion*

Our discussion, largely inspired by the work of Glanzberg, has led to an alternative proposal. I haves suggested, agreeing with Glanzberg, that the idea of faultless disagreement is dubious. But from there, an alternative route opens. Consider the options in relation to a statement of taste, of the form A is Φ. We noted that the 1st order options are simple. We can have naive non-dogmatist experiencer who simply claims that A is Φ and that's it. On meta-level such an experiencer is simply agnostic about further matters: is A Φ for other people, who is right about it, and so on. One alternative, a bit more reflective stance is the dogmatist one: If you don't agree, you just don't know about A being Φ. I think people who do sincerely debate the issues are honest dogmatists, who naively believe they are objectively right. The other option is the tolerant, liberal one: "A is Φ; for me, I mean. How do *you* find it?" On the meta-level, dogmatic disagreement goes well with value-absolutism, entailing that one of the parties is simply wrong, and with relativism. If one is not dogmatist about taste predicates, one should accept that dogmatist is simply wrong; no faultlessness is present. The liberal stance goes well with contextualism. If one is liberal there is no

"Matisse *may* yet be a better painter than Picasso." (italics mine); "Matisse and Picasso", May 19, 1920, available at https://newrepublic.com/article/91909/matisse-and-picasso.

deep disagreement. So, the idea of faultless disagreement is a myth. In this case, liberalism is wiser than dogmatism.

But note that language is open to all possibilities. The language of taste attitudes is compatible with all three first-order stances: with naive non-dogmatism, with dogmatism and with tolerant liberalism. Particular uses of language can be classified along second-order options, as agnostic, absolutistic, relativistic and contextualist. But the whole business is linguistically correct, syntactically, semantically and pragmatically, so I am doubtful that there is a single correct reading of the use of taste predicates and the like. Our agnostic is linguistically in the clear. The absolutist does not reform language, she is into postulating objective value-properties in the world. The relativist is not making a linguistic mistake; her mistake might be rather about the reality of values. Finally, the contextualist is in clear, as far as language alone is concerned; her description fits the liberal usage perfectly, she may only have problems in theoretical accounting for other options, but not with mischaracterizing language as used by the tolerant liberal.

This alternative route might be worth exploring. And to conclude with a hedonic taste statement, it was great fun reading Glanzberg's paper, discussing it with him in Dubrovnik, and thinking about it afterwards!

References

Barthes, R. 1989. *Sade, Fourier, Loyola.* Trans. Richard Miller. Berkeley: University of California Press.

Devitt, M. 2013. "Three Methodological Flaws of Linguistic Pragmatism." In C. Penco and F. Domaneschi (eds.). *What is Said and What is Not: The Semantics/Pragmatics Interface.* Stanford: CSLI Publications: 285–300.

García-Carpintero, M. and Kölbel, M. (eds.). 2008. *Relative Truth.* Oxford: Oxford University Press.

Glanzberg, M. 2005. "Focus: A Case Study on the Semantics/ Pragmatics Boundary." In Z. G. Szabó (ed.). *Semantics vs. Pragmatics.* Oxford: Oxford University Press: 72–110.

Glanzberg, M. 2007. "Context, Content, and Relativism." *Philosophical Studies* 136 (1): 1–29.

Glanzberg, M. 2016. "Not All Contextual Parameters Are Alike." (Preliminary draft of June 12).

Hume, D. 1978. *A Treatise of Human Nature.* Reprint. Ed. Lewis Amherst Selby-Bigge. Oxford: Clarendon Press.

Hume, D. 1983. *An Enquiry Concerning the Principles of Morals.* Reprint. Cambridge: Hackett.

Kaplan, D. 1989. "Demonstratives". In J. Almog, J. Perry and H. Wettstein (eds.). *Themes from Kaplan.* Oxford: Oxford University Press: 481–563.

Kennedy, C. 2007. "Vagueness and grammar: The semantics of relative and absolute gradable adjectives." *Linguistics and Philosophy* 30: 1–45.

Kivy, P. 2001. "Foreword." In N. Carroll. *Beyond Aesthetics.* Cambridge: Cambridge University Press.

Kivy, P. 2015. *De gustibus: Arguing About Taste and Why We Do It*. Oxford: Oxford University Press.

Kölbel, M. 2011. *Objectivity, Relativism and Context Dependence*. Hagen: Fernuniversität in Hagen, Institut für Philosophie: Ch. 5. Extending Kaplan's Framework: Relativism.

Lasersohn, P. 2005. "Context dependence, disagreement, and predicates of personal taste." *Linguistics and Philosophy* 28: 643–686.

Lasersohn, P. 2017. *Subjectivity and Perspective in Truth-Theoretic Semantics*. Oxford: Oxford University Press.

Levinson, J. 2016. *Aesthetic pursuits*. Oxford: Oxford University Press.

Recanati, F. 2010. *Truth-Conditional Pragmatics*. Oxford: Oxford University Press.

Stainton, J. 2005. "In Defense of Non-Sentential Assertion." In Z. Gendler Szabo (ed.). *Semantics vs. Pragmatics*. Oxford: Clarendon Press: 383–457.

Stainton J. 2006. *Words and Thoughts Subsentences, Ellipsis, and the Philosophy of Language*. Oxford: Clarendon Press.

Wright, C. 2008. "Relativism about Truth Itself: Haphazard Thoughts about the Very Idea." In García-Carpintero and Kölbel (2008): 157–186.

Young, J. O. (ed.). 2017. *Semantics of Aesthetic Judgements*. Oxford: Oxford University Press.

Croatian Journal of Philosophy
Vol. XVIII, No. 54, 2018

On the Measurability of Measurement Standards

PHIL MAGUIRE and REBECCA MAGUIRE
National University of Ireland, Maynooth, Ireland

Pollock (2004) argues in favour of Wittgenstein's (1953) claim that the standard metre bar in Paris has no metric length: Because the standard retains a special status in the system of measurement, it cannot be applied to itself. However, we argue that Pollock is mistaken regarding the feature of the standard metre which supports its special status. While the unit markings were arbitrarily designated, the constitution, preservation and application of the bar have been scientifically developed to optimize stability, and hence predictive accuracy. We argue that it is the 'hard to improve' quality of stability that supports the standard's value in measurement, not any of its arbitrary features. And because the special status of the prototype is tied to its ability to meet this external criterion, the possibility always exists of identifying an alternative, more stable, standard, thereby allowing the original standard to be measured.

Keywords: Measurement standard, stability, accuracy, prediction.

1. *Introduction*

Wittgenstein (1953: 29, §50) makes the following claim:

> "There is *one* thing of which one can say neither that it is 1 metre long, nor that it is not 1 metre long, and that is the standard metre in Paris. – But this is, of course, not to ascribe any extraordinary property to it, but only to mark its peculiar role in the language-game of measuring with a metre-rule."

Pollock (2004) argues that this claim is correct. Because the prototype metre has a special status in the metric system, it cannot be measured within that system. His view is that there is no fundamental unit of length beyond the prototype metre. It is not the case that bar was selected because it happens to match an *a priori* concept of the metre. Rather, the bar is the essence of the metre and measuring metric length does not make sense without it.

According to Pollock (2004: 153), “measurement consists in nothing more than the comparison of the object of measurement with some (arbitrarily chosen) standard”. In other words, the value of measurement comes not from some intrinsically meaningful process of evaluation, but from a process of comparison enabled by an arbitrary standard. When we ask “how long is that object?” we are not seeking information about its true length, whatever that might mean. After all, we can see exactly how long the object is. What we want to know is how its length compares with that of other objects, a comparison process which requires some arbitrarily selected standard for quantifying length, be it metres, feet, hands or fingers. Pollock explains that “measurement simply consists in determining the ratio of one object’s length to the length of some standard”. Wittgenstein (1953: 103, §279) makes a related observation, highlighting the meaninglessness of measurement without comparison:

> “Imagine someone saying: ‘But I know how tall I am!’ and laying his hand on top of his head to prove it.”

Taking another example, somebody might step outside on a warm day and exclaim “I wonder how hot it is?” Clearly they can feel how hot the air is, it’s touching their skin. But measurement is not about providing independent, theory-free descriptions of phenomena, it’s about relating things together. What this person wants to know is how the air temperature today compares with that of previous days. In sum, the utility of measurement comes about, not from the result of the measurement itself, but from the comparisons it enables.

Pollock (2004: 152) argues that this feature of measurement, a system for comparison rather than description, has been lost on the majority of philosophers: “...philosophers simply [do] not understand the concept of measurement”. Salmon (1986: 210) is proposed as epitomising this confusion:

> “...if the reference-fixer does not know how long S is, he cannot know, and cannot even discover how long anything is. Measuring an object’s length using S only tells him the ratio of that object’s length to the length of S.”

Also (208):

> “If one knows only that the length of the first is n times that of the second without knowing how long the second object is, one knows only the proportion between the lengths of the two objects without knowing how long either object is.”

Pollock (2004: 149) states that Salmon here demonstrates a failure “to understand the very concept of measurement, as well as what it means to know the length of something.” There is nothing to the act of measurement beyond expressing relationships between objects. There is no absolute scale of measurement, no apodictic system for quantifying unit length. And with no natural unit, there is no concept of measurement beyond an arbitrarily selected standard being used to express length ratios. For Pollock, that’s all there is to measurement.

The question "how long is that object?" presupposes a system of measurement involving a standard of comparison. Because the prototype metre is a necessary condition for the existence of the metric system, the question "how long is the standard metre?" is not a proper question. The standard is a criterion for measuring in the metric system, and it makes no sense to apply a criterion to itself (Pollock 2004: 155). The description of the prototype metre as "one metre long" only functions as a name or label, not as the description of a measurable length of the object. Pollock (2004) therefore concludes that Wittgenstein is correct: We cannot say that the standard metre is a metre in length, or that it is not a metre in length.

In summary, Pollock's (2004) argument hinges on first, the idea that the standard metre is identified arbitrarily, and second, that it retains a special status in the metric system and so cannot be applied to itself. This article investigates whether these two assumptions are valid. In brief, we will argue the following: Pollock is correct in assuming that there is nothing to the act of measurement beyond expressing relationships between objects. However, he is wrong in assuming that this implies that measurement standards are selected arbitrarily (and hence immune to being measured themselves). Not all systems are equally capable of expressing relationships in a useful way. Specifically, standards that feature the property of *stability* are better, because they enable superior predictions. Because measurement standards are obligated to meet the external property of stability, they are susceptible to being improved upon, and thus open to being measured themselves. Although temporarily enjoying dominant status, working standards do not have immunity to being overthrown in the game of measurement. Accordingly, Wittgenstein's statement about the metre bar having no measurable length value must be wrong.

2. *A Brief History of Length*

Before examining Pollock's (2004) argument, we provide a brief overview of the history of the metre and the standard metre bar in Paris.

Measurement standards in medieval Europe varied widely between different jurisdictions, which were often little more than single market towns. The French revolution in 1789 provided the motivation to abolish the multitude of length measures associated with the *ancient régime* and replace them with a new decimal system based on a universal and easily replicable standard (Crease 2011).

The new movement towards standardization provoked much debate as to which environmental property could provide a globally recognizable standard. One proposal was to use the length of a "seconds pendulum", that is, a pendulum which swings through a half-period in exactly one second. However, it was soon discovered that the length of such a pendulum actually varies from place to place. For example, the French astronomer Jean Richer demonstrated a 0.3% difference in this length when calibrated in Cayenne (in French Guiana) versus Paris (Crease 2011).

In light of this, the commission for measurement reform eventually came to the decision that the new unit of length should be equal to one ten-millionth of the distance from the North Pole to the Equator, when measured along the meridian passing through Paris. This was a concept expressed in a single sentence that everybody on earth could agree on. During the surveying process, the commission ordered the production of a series of platinum bars based on preliminary calculations. Following the survey's completion, the bar with length closest to the meridional definition was identified. This bar, which subsequently became known as the *"mètre des archives"* was placed in the National Archives on June 22nd 1799 (Wikipedia "History of the metric system", 2018).

The simple meridional definition had been intended to ensure international reproducibility. In practice, however, nobody was in a rush to replicate a survey of the distance between the Equator and North Pole. The definition was so impractical to verify that it became irrelevant, being replaced instead by artefact standards. When it was later established that the circumference quadrant was actually 10,019km, as opposed to 10,000km, this had no bearing on the use of the metre. The use of artefacts was already providing a de facto standard, unconnected and arbitrary relative to any other worldly definition.

Countries adopting the metre as a legal measure during the 19th century purchased standard metre bars with which to calibrate their own national standards. These, however, were prone to wearing down with use. Because different standard bars in different countries were being worn down at different rates, there was no mechanism for verifying whether everybody was adhering to the same standard. In light of these difficulties, an international treaty, known as the Metre Convention, was signed in Paris on 20th May 1875. An organisation known as the Bureau International des Poids et Mesures (BIPM) was established in Sèvres, just outside Paris. This organization was entrusted with the responsibility of conserving prototypes and carrying out regular comparisons between different national standards, so as to ensure international consensus.

The BIPM set about creating a new state of the art international prototype metre, accompanied by a set of copies earmarked for international distribution. These bars were made of a special alloy, consisting of 90% platinum and 10% iridium, making them significantly harder than pure platinum. They were also fashioned in the shape of an X, thus minimizing the effects of torsional strain during length comparisons.

One of these bars was "sanctioned" to be identical in length to the *mètre des archives* on September 28th 1889, during the first meeting of the Conférence Générale des Poids et Mesures (CGPM). Following this moment of consecration, the new bar became the international prototype metre, and the old 1799 bar began to fluctuate in length.

In 1960, at the 11th CGPM, a new definition of the metre was agreed, based on wavelengths of radiation from the krypton-86 atom. In 1983, at the 17th CGPM meeting, the metre was redefined again in terms of the distance travelled by light in a vacuum per second.

For the purpose of analysing the validity of Wittgenstein's original claim and Pollock's (2004) defence of it, we will initially consider the role of the 1889 metre bar as an active standard, as it was in 1953 when Wittgenstein's comments were first published.

3. *Length versus Unit*

Pollock (2004) repeatedly emphasises that the selection of the standard is arbitrary, meaning that it is completely self-sufficient and has no connection with any external phenomena: The "standard is arbitrarily chosen and agreed upon by the community. Only practical considerations bar us from using anything at all as a standard" (154). Also: "This arbitrary nature of standards of measurement seems to be lost on many philosophers" (154).

Intuitively, the selection of prototypes by the BIPM does not seem arbitrary. For example, the prototype kilogram is deliberately forged of platinum-iridium alloy, an inert metal with very high density (to negate a buoyancy effect), extreme resistance to oxidation, low magnetic susceptibility and high resistance to contamination and wear. In addition, the artefact is carefully isolated under multiple nested bell jars and subject to periodic cleaning with ether and ethanol followed by steaming with bi-distilled water (Wikipedia "Kilogram", 2018). When Pollock describes the prototype metre as arbitrary, he is referring, not to its material, preservation and application, but to the markings on that bar which designate one metre. It is the designation of a *unit* that is arbitrary.

But what is the size of that unit? We propose that the size of the unit does not exist independently of the medium of the platinum-iridium bar onto which it is inscribed. The bar does the work of preserving the size of the unit, rendering the concepts of 'unit' and 'standard' inextricable. Asserting that the unit is arbitrary is therefore meaningless: there *is* no unit that can be addressed independently of its embodiment by the metre bar itself.

All measurement units depend on an underlying standard which embodies their size. For example, the Imperial and metric systems were originally associated with different processes for realizing their respective units. However, by 1964, the definition of the inch was *tied* to that of the metre, meaning that both units serve as different labels for describing measurements in the same fundamental system. To turn a measurement from centimetres to inches, one simply divides by 2.54. Although this ratio is one that arose by historical chance, it has no measurement value in itself, serving merely as a cosmetic treatment of an underlying measurement result. The value of both the Imperial and metric systems lies in the embodiment of unit length by a standard.

Pollock (2004) misses the idea that the size of a unit must be realized by some sophisticated practice, believing instead that the concept of objective length is universally appreciated following its ‘discovery’: “Although we discovered the concepts of length (and mass) we invented the concept of a metre for our own convenience; as a means of making judgments about length, which we could record and/or communicate to others” (154). Thus, for Pollock, the aspect of the prototype metre that gives it its special status in the metric system is not its role in enabling reliable judgments about length, but merely its role in designating a unit of measurement. In the following section we argue that Pollock’s attitude overlooks a crucial property of measurement standards, namely that of *stability*.

4. *What Makes the Standard Special?*

Let’s imagine what would happen if the only role of the metre bar was to designate a unit of measurement, as Pollock (2004) assumes, without any regard to realizing the size of that unit.

Under this scenario units of length could be perfectly replicated and maintained by any measured object. For example, I could take a wooden stick and mark on it exactly the same unit lengths as exist on the prototype metre bar. In Pollock’s world the stick functions just as well as the original standard. Indeed, every act of measurement is equivalent to forging a new standard. Once my desk is identified as having some particular length value, it too becomes part of the standard, and, following the assumptions inherent to Pollock’s view, ceases to have measurable length because of its new special role in the system. The original standard is not special anymore. We can no longer cling to Wittgenstein’s statement that only *one* thing has no measurable length value: Every object which is measured becomes just as good at realizing length as the original standard, hence losing its property of measurable length. Pollock doesn’t care what object the markings are made on. After all, the choice is arbitrary.

In practice, measurement does not work like this. The standard encompasses, not just the physical bar, but a whole set of procedures for handling, comparing and making copies of the bar, as well as the background knowledge and assumptions involved in those procedures. For example, in 1927 the defined *mise en pratique* of the prototype metre was altered, without affecting the prototype artefact, or its unit markings. At the 7th CGPM it was clarified that any measurement of the bar should now be “subject to standard atmospheric pressure, with the prototype supported on two cylinders of at least one centimetre diameter, symmetrically placed in the same horizontal plane at a distance of 571 mm from each other” (BIPM 1928: 49). The preservation of a given measurement standard resides in the understanding of its *mise en pratique* by active practitioners; used improperly the metre bar might prove no more useful in measurement than a metre stick.

If we accept that some artefacts and procedures enable superior judgements about length (e.g. a platinum-iridium bar, when used in appropriate manner, makes a better standard than a stick) then we are admitting the existence of some external criterion that standards are intended to meet.

Consider, for example, a fanatical dictator who issues a diktat defining the length of his beard as the new standard for measurement. Relying on this unstable beard length might cause bridges to fall down, buildings to collapse, and ships to sink.

Is this a problem? If we maintain that a standard of length can be selected arbitrarily, then it has no obligations to achieve anything. It is only relative to the external goal-directed expectation that measurement standards should keep bridges up and ships afloat that we can describe the beard-length standard as wanting. In sum, an arbitrary standard, without external connection to any practical function, does not support the property that we intuitively understand as length.

Danjon (1929) highlights the difficulty of interpreting the ephemeris time standard (using the position of the sun, moon, planets and stars) as a fiat with no external obligations:

> …Although Newton's law has been saved, it is experiencing a quite extraordinary adventure: henceforth called upon to gauge the passage of time, it becomes in part unverifiable and ceases to be what could strictly be termed a law…Since we would ask these laws to provide a measure for the passage of time, we could no longer subject them to experimental control without entering into a vicious circle. (Danjon 1929)

Consistent with Danjon's critique, Chang (2001, 2004, 2007), van Frassen (2008, 2009) and Tal (2011, 2012, 2013, 2016) all reject the traditional view of arbitrary, apodictic definitions at the heart of measurement. Acknowledging the real-world application of measurement, they recognise the role of a 'hard-to-isolate' external criterion, supporting goal-directed activities, as being at the heart of the practice.

5. *What is Measurement For?*

To properly understand the role of the prototype metre in the metric system we need to consider what measurement is for in the first place.

Intuitively, people make measurements because measurements are valuable. But what is it about measurements that makes them useful? Tal (2012, 2013, 2016) proposes that the goal of prediction lies at the heart of measurement, insofar as measurement accuracy, and hence the calibration of scientific instruments, is defined in terms of predictive accuracy. When I measure the length of my desk I am effectively making a prediction about what will happen when it interacts with other measured objects (e.g. will my desk fit through that door?) Even if we imagine cases where measurement is carried out for its own sake, without any expectations for prediction, the concept of reliable relationships still applies. For example, somebody who measures how fast they run around a race track expects those timings to enable comparisons involving other runners, suggesting who would win a hypothetical race between them. In order to be of value, a measurement system must provide reliable information about the relationships between measured phenomena, information which enables accurate predictions.

Tal's (2012) goal-based view stands in contrast to the widespread supposition that measurement and prediction are distinct epistemic activities. He argues that traditional accounts of measurement have overlooked its practical role in prediction, ignoring the key associated concepts of uncertainty, reliability and inference. For example, theorists such as Campbell, Stevens and Suppes "took 'measurement' to be synonymous with either 'number assignment' or 'scale construction', and neglected the 'applied' aspects of measurement such as accuracy, precision, error, uncertainty, and calibration" Tal (2013: 1164). In practice, measurement outcomes are obtained from instrumental readings by a chain of inferences, and the inferences drawn depend on the particular theoretical and statistical assumptions associated with the measurement apparatus. According to Tal (2013: 1165), "this way of viewing measurement raises a host of representational questions that have been either neglected or only partially addressed by traditional accounts".

The idea that measurement might be goal-directed raises the issue of how a theoretical quantification could be coordinated with empirical measurement. The issue here is that the empirical adequacy of a given theory and the reliability of a related measurement process appear to depend on each other in a circular fashion (Tal 2013: 1160). For example, in order to establish a theory of weight, it is necessary to test the predictions of that theory, a task which itself requires a reliable method of measuring weights. Conversely, testing the reliability of such measurements presupposes existing theoretical knowledge about weight against which it can be calibrated (Tal 2013: 1160).

The traditional philosophical approach to this problem, which Pollock (2004) espouses, has been to assume that coordination is achieved, and circularity avoided, by establishing apodictic definitions for quantification, which are arbitrary, self-supporting and internally complete. These definitions are assumed to be "analytic statements that require no empirical testing" (Tal 2013: 1160), thus severing the link between measurement and any external goal-directed outcome, such as prediction. For example, Ernst Mach noted that different types of fluid expand at different nonlinearly related rates when heated and concluded that there can be no fact of the matter as to which fluid expands most uniformly, since the very notion of equality among temperature intervals has itself no determinate application prior to a conventional choice of standard thermometric fluid with which to establish it (Tal 2013: 1161). The eventual choice of standard, for Mach, was a conventional one. Poincaré similarly argued that the processes scientists use to mark equal time durations (e.g. pendulum swings) are chosen for the sake of convenience (Tal 2013: 1161).

Pollock (2004: 155) echoes a similar conventionalist sentiment when he insists that "we simply chose a length that we found convenient and *called* it a metre. That is all there is to choosing a standard of measurement." However, though the examples noted by Mach and Poincaré seem, at first blush, to indicate arbitrariness at the heart of measurement, this arbitrariness results from pushing measurement beyond the existing limits of science and technology, thereby exhausting justification. The arbitrary decision here is to choose between several highly sophisticated systems, each of which does so well at measuring that their various merits are hard to distinguish.

For instance, while Mach and Poincaré recognized that choices of coordinative principles are often constrained by considerations of simplicity and convenience, they were not suggesting that these choices are completely arbitrary, but rather that working standards are selected because they are "good enough" to provide useful practical reference (see Galison 2003), an attitude subsequently adopted by the BIPM.

6. *Stability*

Metrology is the science of measurement and standardization, carried out by metrologists, who are experts in highly reliable measurement. Despite the fact that it is an independent discipline with its own journals and controversies, the methods and tools of metrology have received little attention from philosophers (Tal 2011: 1083). A central philosophical question in metrology is how the process of standardization works. What exactly is it that metrologists are doing to develop and maintain accurate standards of measurement? How are these methods justified from an epistemic perspective and how do they resolve the apparent circularity of theoretical quantification and empirical measurement?

Chang (2001, 2004, 2007) and van Fraassen (2008, 2009) argue that the apparent circularity is not vicious. According to their view, constructing a quantity concept and standardizing its measurement are co-dependent, iterative tasks. With each iteration the quantity concept is re-coordinated to a more stable set of standards, which allows theoretical predictions to be tested more precisely, facilitating the subsequent development of standards, and so forth (Tal 2013: 1162). This corresponds with the BIPM's view of their own standards, which are not intended as absolute but rather based on a *'mise en pratique'*, that is, a set of instructions allowing the unit to be realized in practice with the highest level of accuracy. The difference between this view and the traditional philosophical approach is that it does not seek to resolve circularity through absolutism. Rather, it treats the standard as a working realization of an external criterion known in metrology as ‘stability’.

Stability refers to the tendency of an apparatus to produce the ‘same’ measurement outcome over repeated runs, as well as replicating the outcomes of similar instruments around the globe. What this means in practice is that discerning any predictable fluctuation in a standard should be as hard as possible; the standard should be as uncorrelated as possible with any changes in the environment. This is the external criterion that measurement standards are designed to meet. Under the guidance of the BIPM, a worldwide network of metrological institutions is responsible for comparing, adjusting, maintaining, disseminating and refining stable standards (Tal 2016: 297).

One of the notable successes of these institutions is the standard measure of time used in almost every scientific context, known as Coordinated Universal Time (UTC) (Tal 2016: 297). UTC is regarded as overwhelmingly stable insofar as a variety of standardization labs around the world manage to closely reproduce it on an ongoing basis. Standardization can be regarded as a process for ensuring independent agreement: Despite being displaced in space and time, and having no causal interaction with each other, the labs can produce results which agree with each other. In other words, they are able to make highly accurate predictions about the measurements that other labs will report each day. Metrologists labour relentlessly to identify standards that support greater predictive accuracy. If standards were chosen arbitrarily, as Pollock (2004) maintains, the world would have no need for metrologists.

As regards the prototype metre bar, its value comes, not from those features which have been selected arbitrarily, but from those which have been carefully calibrated to maximise stability. The standard reflects the realisation of centuries of accumulated theoretical and technological efforts, involving the identification of materials that best support predictive accuracy under varying conditions. Contrary to Pollock’s understanding, the metre bar’s utility is not related to its ability to *designate* a unit of length. After all, anyone could just as well hold their two fingers in the air, refer to the distance between them, and say “this is the length of a metre”. Designating an arbitrary distance is easy, but to be rendered useful, the size of that unit must be preserved by some stable standard. The utility that the prototype metre bar provides lies in its capacity to maintain and replicate that designated distance.

7. *Measuring the Standard*

In sum, measurement and stabilization are one and the same concept. To measure a property such as length is to stabilize it relative to a standard which can reliably preserve that property, thus enabling accurate predictions to be made. Stability is the backbone of measurement utility, and working standards merely approach that ideal without ever realizing it completely.

We return now to the original question of whether the prototype metre bar has measurable length. We have argued that what makes measurement standards valuable is their capacity to enable reliable judgements about length, and hence support accurate predictions. If a standard has been designed to meet an external goal-based criterion, this opens up the possibility of improving the standard and replacing it with a more stable version, thus allowing the original system to be measured. Because stability relates to external events and relationships, no standard can ever represent the final word on stability. As soon as we identify measurement value as being related to stability, we recognize that working standards provide a useful, yet incomplete representation of the concept of measurement.

A continuing trend in metrology is to eliminate as many as possible of the artefact standards, and instead define practical units of measurement in terms of fundamental physical constants. As of writing, the only remaining artefact standard is the International Prototype Kilo (IPK), shortly to be replaced, like that of the other BIPM base units, by a definition entirely based on fundamental constants.

For example, the metre bar prototype was officially superseded in 1960, at the 11th CGPM, when a new definition was agreed based purely on a universally replicable *mise en pratique*. Specifically, the metre was redefined as equal to 1,650,763.73 wavelengths in a vacuum of the radiation corresponding to the transition between the levels $2p^{10}$ and $5d^{5}$ of the krypton-86 atom. This new definition democratised and diversified the materialization of standards, by allowing anyone with the appropriate lab equipment to realize the metre for themselves. Increasing levels of scientific sophistication and greater levels of shared practical knowledge between metrologists have obviated the need for a remaining link to a localised artefact.

It should be noted that the shift from artefact to decentralized standards has not changed the practicalities of metrology any more than the de jure abandonment of the gold standard in 1976 changed the nature of international economics. In practice, the Parisian artefact standards were rarely consulted. The *only* comparison of national standards with the international prototype was carried out over a 15 year period between 1921 and 1936, revealing a variability of around 0.2 μm (Nelson 1981). Like the gold standard, the role of artefact standards was chiefly to shore up confidence in the system as a whole. As scientific knowledge became more widespread, sophisticated and interconnected, this role was no longer necessary.

The definitions of decentralized standards, just like artefact standards, involve an arbitrary component which is needed to establish a convenient unit quantity. For example, in the krypton-86 standard, the value “1,650,763.73” was selected so as to ensure historical continuity with the preceding definition. The number is arbitrary, insofar as any other number would work just as well. However, as previously argued, the number by itself does not provide utility. Instead, it's the stability of krypton radiation wavelengths that supports reliable judgements about length.

The issue of standards being vulnerable to measurement applies just as equally to decentralized standards as it does to localized artefact standards. In order to maintain their status, measurement standards are obligated to deliver reliable judgements, and to support accurate predictions. When competitors can gain an advantage using an alternative system, a working standard immediately loses its status.

The new BIPM base units, which tie base units to fundamental constants, state ideal conditions that cannot be realized by a material object or process, only by an abstract entity, these conditions can be approached more and more closely in practice, yet never perfectly realized (e.g. achieving a perfect vacuum to measure the speed of light). Accordingly, the realization of standards is left entirely open and prone to change when metrologists discover new physical principles that make it possible to materialize the unit with greater stability than before.

Just as with artefact standards, the incomplete understanding of stability leaves decentralized standards perennially vulnerable to refinement.

8. *Conclusion*

Pollock's (2004) argument begins promisingly, with the observation that the utility of measurement stems from its capacity to support comparisons, and not from providing absolute, theory-free descriptions. However, he makes a critical error by falling back into the trap of absolutism, assuming that the concept of length is 'objectively' known, as opposed to something whose practical realization we must work relentlessly towards.

The prototype metre in Paris was selected by metrologists as a useful working standard because it did a good job. It was never intended as the absolute, inviolable definition of the metre. Pollock's (2004) arguments regarding the irreproachable role of the metre standard are directly undermined by the BIPM's 1960 declaration from the 11th CGPM, according to which "the international prototype does not define the metre with an accuracy adequate for the present needs of metrology" (Tal 2011: 1082–1083). If the metre bar was really the foundation of measurement, how could its accuracy ever be found lacking?

Pollock (2004) overlooks the crucial idea that measurement is a goal-directed activity based on clear external objectives, and thus open to continuing refinement. When a system asserts its own supremacy, it severs any ties to delivering in practice, and the system ceases to have utility. For instance, any measurement standard which is beyond reproach, such as the dictator's beard, cannot measure at all, because it is freed of any responsibility to provide practical results in the real world. To be useful, a measurement standard must hold the potential to be found lacking—to be measurable—by some alternative system which delivers superior results in practice. Thus, while current measurement standards do a great job, they do not completely define what we demand of measurement.

For example, in 1988, the International Prototype Kilogram (IPK), which continues to serve as the standard for mass, was removed from its vault in Paris. It was found that the mass of the prototype had drifted downwards relative to the set of national copies distributed globally in 1884, at a rate of change of about 0.5 parts per billion per year (Crease, 2011). By definition, the prototype has no measured value, and hence no measured error. From this frame of reference, the copies around the world are gaining mass. However, because that is a clearly counterproductive interpretation, the BIPM 'inferred' that the prototype must be unstable and somehow losing mass, thus making an implicit comparison of the mass of the IPK to some more stable reference frame.

In conclusion, Wittgenstein's original claim regarding the measurability of the prototype metre must be mistaken. The prototype holds its status as a standard, not because it has been arbitrarily singled out as having a special role in some language game, but because it delivers results in practice which are hard to beat. Measurement standards should thus be interpreted as well-established recommendations for how to achieve the best possible measurement results given the current state of technology. As soon as we succumb to the assumption that standards somehow encapsulate the foundations of measurement itself, and are thus immune to reproach, we cease to be engaged in measurement.

References

BIPM 1928. Comptes Rendus de la 7e CGPM (1927).

Chang, H. 2001. "Spirit, air, and quicksilver: The search for the 'real' scale of temperature." *Historical Studies in the Physical and Biological Sciences* 31 (2): 249–284.

Chang, H. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Chang, H. 2007. "Scientific progress: Beyond foundationalism and coherentism." *Royal Institute of Philosophy Supplement* 61: 1–20.

Crease, R. P. 2011. *World in the balance: the historic quest for an absolute system of measurement*. New York: W. W. Norton & Company.

Danjon, A. 1929. "Le temps, sa definition pratique, sa mesure." *L'astronomie* XLIII: 13–22.

Galison, P. 2003. *Einstein's Clocks, Poincaré's Maps: Empires of Tim*e. New York: W. W. Norton & Company.

Nelson, R. A. 1981. "Foundations of the interational system of units (SI)." *The Physics Teacher* 596–613.

Pollock, W. J. 2004. "Wittgenstein on the standard metre." *Philosophical Investigations* 27 (2): 148–157.

Salmon, N. U. 1986. *Frege's puzzle.* Atascadero: Ridgeview Publishing Company

Tal, E. 2011. "How accurate is the standard second?" *Philosophy of Science* 78 (5): 1082–1096.

Tal, E. 2012. *The Epistemology of Measurement: A Model-Based Account.* PhD Thesis, University of Toronto.

Tal, E. 2013. "Old and new problems in philosophy of measurement." *Philosophy Compass* 8 (12): 1159–1173.

Tal, E. 2016. "Making time: A study in the epistemology of measurement." *British Journal for the Philosophy of Science* 67 (1): 297–335

van Fraassen, B. 2008. *Scientific Representation: Paradoxes of Perspective.* Oxford: Oxford University Press.

van Fraassen, B. 2009. "The perils of Perrin, in the hands of philosophers." *Philosophical Studies* 143: 5–24.

Wikipedia "History of the metric system" (2018, February 9). In *Wikipedia, The Free Encyclopedia.* Retrieved 17:49, February 9, 2018, from https://en.wikipedia.org/w/index.php?title=History_of_the_metric_system&oldid=823583734

Wikipedia "Kilogram". (2018, February 9). In *Wikipedia, The Free Encyclopedia.* Retrieved 17:49, February 9, 2018, from https://en.wikipedia.org/w/index.php?title=Kilogram&oldid=824750032

Wittgenstein, L. 1953. *Philosophical Investigations.* G. E. M. Anscombe and R. Rhees (eds.). G. E. M. Anscombe (trans.). Oxford: Blackwell.

Croatian Journal of Philosophy
Vol. XVIII, No. 54, 2018

Evolution and Ethics: No Streetian Debunking of Moral Realism

FRANK HOFMANN
University of Luxembourg, Luxembourg

This paper is concerned with the reconstruction of a core argument that can be extracted from Street's 'Darwinian Dilemma' and that is intended to 'debunk' moral realism by appeal to evolution. The argument, which is best taken to have the form of an undermining defeater argument, fails, I argue. A simple, first formulation is rejected as a non sequitur, due to not distinguishing between the evolutionary process that influences moral attitudes and the cognitive system generating moral attitudes. Reformulations that respect the distinction and that could make the argument valid, however, bring in an implausible premise about an implication from evolutionary influence to unreliability. Crucially, perception provides a counterexample, and the fitness contribution of reliably accurate representation has to be taken into account. Then the moral realist can explain why and how evolution indirectly cares for the truth of moral attitudes. The one and only condition that has to be satisfied in order for this explanation to work is the sufficient epistemic accessibility of moral facts. As long as the moral facts are sufficiently reliably representable, one can see how evolution could favor getting it right about the moral facts. Interestingly, apart from this epistemic constraint no further constraint and, in particular, no objectivity constraint on what the moral facts have to be like can be derived. Thus, the only problem for the moral realist is to make good on epistemic access to moral facts—an old problem, not a new one.

Keywords: Evolution, ethics, debunking, moral realism, reactive attitudes.

1. *Introduction*

Discrediting a view or set of belief-like attitudes that aspire to truth by so-called 'debunking' has become quite popular.[1] An especially interest-

[1] See, for example, Kahane (2011) for a list of examples and some general discussion of debunking arguments.

ing case is the case of moral attitudes and evolution. The debunking argument attacks our moral attitudes (or an important part of them)[2] by pointing out that the force that has shaped these moral attitudes—evolution—is blind to issues of truth and only cares about fitness and survival.[3] Take 'altruistic attitudes' as the core of our moral attitudes. An evolutionary explanation of these attitudes can be given by reference to the fitness-conduciveness of certain kinds of cooperative behavior towards one's kin or group. If these altruistic attitudes are understood realistically, we run into a problem: they might be true, but only accidentally so, since the relevant evolutionary forces did not care about truth. Or so the debunking line of thought runs.

Not too long ago, Sharon Street has presented an evolutionary debunking argument against moral realism, called the 'Darwinian Dilemma'.[4] So she is turning the argument around, targeting the realistic assumption that the relevant attitudes aspire to truth and are, at least sometimes, true—what is often called *'moral realism'.* (Whether realism is or should be construed as implying some objectivity condition can initially be left open and will be investigated later.) The argument is clearly of high significance, since it attempts to bring to fall an entire approach in (meta-)ethics, i.e., moral realism. Street's formulation of the argument, however, is burdened with considerations that rather distract from those valuable and interesting points that may constitute a sound argument against moral realism. One could try to disentangle the various claims and streams of thought in her presentation(s) of the argument, which would require an immense amount of careful interpretational work. Here I would like to proceed in a different way, namely, by providing a clear and systematic reconstruction of a core argument that can be extracted from Street's considerations, without any side roads or unnecessary accompaniments. If it turned out that the argument is not really Street's argument, this should not be too worrisome, since it is at least similar and in any case important enough.[5]

[2] I will take the relevant moral attitudes to be beliefs or sufficiently belief-like, since (only) they are aspiring to truth. Any difference should not matter to the argument. Street includes desires (and attitudes of approval and disproval) among the evaluative attitudes (cf. Street 2008: fn. 3). However, these are not really at stake since they are not truth evaluable and, therefore, the issue of reliability does not arise for them, at least on a standard conception of desires.

[3] A particularly succinct statement of this claim—or dogma, indeed—can be found in Burge (2010), in particular, ch. 8. Burge discusses this idea in the context of a naturalistic teleosemantics which conceives of representational functions as biological functions, and tries to argue that it has to fail for exactly that reason. For a convincing criticism of Burge's argument see Graham (2014). For another statement of the claim see, for example, Stich (1990: 62).

[4] The original statement is to be found in Street (2006). Street explains the argument further and, in particular, defends it against the criticism by Copp (2008) in Street (2008). Street takes value realism in general as her target. In order to keep the discussion in reasonable bounds here, I will restrict myself to the moral part of value realism, i.e., to moral realism.

[5] There are several problematic aspects to the exposition of the 'Darwinian Dilemma' as to be found in Street's writings. I only mention a few of them here. It is

In the following I will try to carve out and (re-)formulate what I take to be the heart of the considerations that can be found in Street's writings and that might provide a genuine evolutionary argument against moral realism.[6] Exegetical issues will not be my primary concern, but systematic reconstruction.[7] I will try to show that in the end, the argument fails. The moral realist can tell a plausible story about how evolution cares for the reliability of moral attitudes—for in a nutshell, evolution *indirectly* cares (or, at least, can care) for truth, and there is a plausible story about how this could work. The new part of the story that will be told here appeals to *reactive attitudes* (in Peter Strawson's sense) *as a social mechanism* that explains how the relevant cognitive process could be reliable after all. It provides a new picture of how evolutionary forces could have *indirectly* resulted in reliability. This adds significant support to the idea of indirect truth tracking as a response to the debunking challenge. Interestingly, this makes appeal to other strategies for defending moral realism superfluous (like appeal to the additional epistemological potential of rational reflection, (quasi-)conceptual truth, or a threat of self-defeat within the debunking argument).[8] Fortunately, as will be argued at the end, there will not

questionable whether putting the argument in the form of a dilemma is fortunate. Construing it as a *reductio* of the moral realist assumption, by way of a defeater argument, seems more appropriate, as I will try to show. – Graber has criticized Street for setting things up in the form of a dilemma, too (cf. Graber 2012: 594). Graber suggests that we should take the argument to be an abductive one, i.e., an argument about best explanation. I disagree. Furthermore, Street's formulation also depends on a demand for reasons (for believing in the reliability of the source of our moral attitudes), or even on a demand for *independent* reasons—thus leading into issues of epistemic circularity that are at most implicitly hinted at by Street but by no means discussed explicitly and to a sufficient extent (cf. Street 2008: sc. 6). Externalists will reject such a demand for reasons. A further point of unclarity is the definition of 'moral realism' as it is the target of the argument. Street burdens 'moral realism' (and equally realism about value) with an element of objectivity which is not really required for any kind of moral realism, but only for an *objective* moral realism (cf. Street 2006: 110; Street 2008: 208 and, in particular, fn. 3). If reliability is the primary issue, why is not any kind of moral realism affected?, one can wonder. (Skarsaune presents some good critical observations about Street's characterization of 'moral realism' in Skarsaune 2011: sc. 5.)

[6] Thus, to repeat, I will restrict the discussion to moral realism (and ignore nonmoral value realism). Therein I follow Copp (2008). Street notes that she sees some difficulties with this restriction since it "introduces crucial complexities having to do with morality/reasons internalism" (Street 2008: 209). I have to confess that I fail to see any significant problem with the restriction.

[7] Some useful exegetical work, including important interpretational questions, has been provided by Copp (2008), Enoch (2010), Skarsaune (2011), and Garber (2012).

[8] Brosnan (2011) and FitzPatrick (2015) try to argue that *rational reflection* can lead to reliable belief formation (even if starting with initially false input beliefs). Conceptual truths, or something close enough, have been offered in response to debunking arguments by Cuneo and Shafer-Landau (2014). The charge of self-defeat is discussed in Kyriacou (2016).

be any new significant epistemological costs to the proposed response in defense of moral realism.[9]

2. *Reconstructing a Streetian evolutionary argument (crude version)*

Here is my proposal about how to formulate the initial, crude argument, argument A1:

A1:
(1) Our moral attitudes (MAs) are significantly influenced by an evolutionary process E.
(2) E is not truth-tracking.
———
(3) Our moral attitudes (MAs) are not (epistemically) justified.[10]

Some comments on the premises are in order. Premise (1) is a rough statement that leaves out which moral attitudes exactly are at stake. I have already indicated that we are talking about the 'altruistic core' of our moral attitudes, concerning positive evaluation of cooperation and family support etc. (even at some cost to one's own self-interest). For brevity's sake I will call them the 'MAs'. (As already mentioned, I will treat the MAs as beliefs or sufficiently belief-like to be epistemically evaluable in the relevant way. Moral attitudes that are not truth-evaluable cannot be the target of the argument.) Of course, there is no precise definition of what counts as 'significant influence'. But it seems to be agreed on by all parties in the discussion that the influence is significant, at least for the sake of the argument. After all, it could turn out to be the case (if it has not yet). So let us take premise (1) for granted for the moment and see where it leads. (We will see soon that some reformulation is necessary.)

[9] Similar arguments against various forms of realism (evaluative realism, religious realism, etc.) can be found in the literature. See, for example, Joyce (2008) and Ruse, Wilson (1995). Interesting comparisons of Street's arguments and Plantinga's evolutionary argument against atheism can be found in Crow (2015) and Moon (2016). A very good overview on evolutionary debunking arguments is Vavova (2015). Vavova also presents her own reconstruction of the most promising evolutionary argument which is different from the one I am reconstruing here. She does not recognize the distinction between evolutionary and cognitive processes that I will argue for in the next section, and she does not discuss the criticism of, and reply to, the argument that I will present in section 4. The same is true of Vavova's earlier discussion in Vavova (2014).

[10] This is essentially an instance of the argument schema that Kahane proposes as a general schema for debunking arguments, see Kahane (2011: 111). Kahane discusses Street's argument as fitting into this schema (cf. Kahane 2011: sc. 3). However, Kahane does not present the criticism that I am going to lay out here. Rather, he gets into issues of objectivity and whether the argument overgeneralizes—which are not the problems I am discussing here. Skarsaune (2011) also gets into the issue of what 'independence of our attitudes' means and whether philosophers like Nagel or Parfit hold that moral truths are 'independent' in this or that sense.

Premise (2) is the statement of evolution's blindness to truth. Evolution does not care for truth but only about survival or practical value. This slogan is taken to amount to a lack of truth-tracking. The evolutionary process E is epistemically evaluated as a bad one: it does not push toward truth, it is not truth-conducive.

Why is (3) supposed to follow?—I propose that we should take the argument to have the form of a *defeater argument* or, more precisely, an *undermining-defeater argument.* [11] Its premises provide an undermining defeater against the epistemic justification of the relevant moral attitudes.[12]

Famously, Pollock distinguished between two kinds of defeaters, rebutting and undermining (or undercutting) defeaters.[13] The rebutting defeaters (against the belief that p) are simply reasons for the opposite belief, i.e., belief that non-p. The undermining defeaters consist in reasons against the reliability of the source of one's belief that p. They undermine the source as not being reliable and, thus, as not issuing (*ultima facie*) justified belief (as long as they are not themselves undermined, of course).[14] I propose to take the evolutionary debunking argument as a defeater argument of the undermining sort. The source is taken to be the evolutionary process E and its lack of truth-tracking is taken to undermine E's reliability.

If taken in this way, we can see how one could think that the conclusion (3) follows from the premises. Premise (1) establishes a source of our MAs, and premise (2) discredits it, so the MAs lose the epistemic status of being justified. It is just like when a belief loses epistemic justification if one acknowledges that it has been generated by unreliable wishful thinking, for example.[15]

Prima facie, A1 looks like a sound defeater argument. The history of our moral attitudes seems to be discredited such that their (epistemic) justification is undermined. For what comes from a source that does not track truth could at best be accidentally true—and that is not good enough for (epistemic) justification.[16]

[11] In general, it seems that 'debunking arguments' can be best understood as undermining-defeater arguments. (Perhaps there are some exceptions, but this should be the rule.) Kahane makes the same proposal (cf. Kahane 2011: 105–6). Schafer also takes the heart of Street's considerations to be aiming at an (*a posteriori*) defeater (of an *a priori* entitlement). And he criticizes the attempt for not relying on the idea that the *empirical* facts of evolutionary theory could provide a defeater for a *normative* claim (cf. Schafer 2010).

[12] It is quite clear that our primary concern here is with *doxastic* justification, not with propositional justification or personal justification.

[13] Cf. Pollock (1986) and, for a general overview on evidence, Kelly (2014).

[14] To be more explicit, a belief might be *prima facie* justified, but any undermining defeater cancels its *ultima facie* justification (as long as it is not undermined itself). Cf., for example, Senor (1996).

[15] Of course, the undermining defeater is supposed to be not defeated itself.

[16] In general epistemology there is quite a controversy about how to understand how (undermining) defeaters exactly work, and some epistemological views seem to have problems here. (For some recent discussions of defeaters see, for example,

However, the appearance is misleading. There are at least two big problems with that argument. First, and most importantly, as it stands, argument A1 is a *non sequitur*. And second, and related to that, it is unclear how premise (2) is really to be understood (in the best, charitable way). Let me explain.

Firstly, the argument is a *non sequitur*. This is so since *E is not a cognitive process*. Therefore, the non-truth-tracking of E, stated in premise (2), is simply *not directly relevant* to the epistemic status of justification of our MAs. (It might be indirectly relevant—we will come back to that in due course.) What matters is whether the relevant cognitive process, i.e., the cognitive process of which our MAs are the outcomes, is reliable or not. But since E is not a cognitive process, E's blindness to truth is not directly relevant. (Its indirect relevance would have to be made explicit, and I will try to do so soon.) As the argument A1 stands, it is a *non sequitur*. Premises (1) and (2) do not provide an undermining defeater against the assumption that the relevant cognitive process of which our MAs are the outcome is reliable. An undermining defeater argument has to provide reason against the reliability *of the relevant cognitive process*, since only the relevant cognitive process's reliability is what matters for the epistemic justification of its outcomes. No such reason has been provided so far.[17] This is the decisive shortcoming of argument A1.[18]

Bergmann (2006) and Hofmann (2013).) But the phenomenon is widely acknowledged, and my discussion does not depend on any controversial account of, or assumptions about, undermining defeaters. In particular, it does not depend on whether one goes for a psychologistic or an anti-psychologistic conception of defeaters (i.e., whether one conceives of defeaters as beliefs or the propositions believed).

[17] That cognitive processes are the relevant items when it comes to epistemic justification and defeaters is a common assumption in general epistemology, at least for those epistemologists which admit the relevance of (certain aspects of) the history of a belief (cf., for example, Goldman 1979). If one opts for an a-historical, 'current time-slice' epistemology, such as, for example, Pryor's dogmatism, then no historical processes are directly relevant to the status of epistemic justification. Then, however, the prospects for an evolutionary undermining defeater argument are even dimmer. So the assumption that the history of an attitude matters is granted for the sake of the argument, and not a substantial assumption that I am making. But if the history is relevant, we have to be clear about which parts or aspects of the history of a belief are relevant. There is of course some significant controversy about the individuation of cognitive processes, whether they are more narrowly or more broadly construed and whether they are to be construed individually or socially. The generality problem is a major topic here. (Cf., for example, Goldman 2008. For a social individuation see Goldberg 2010.) But some limits are highly plausible and commonly accepted. In particular, it is quite clear that a cognitive process does not extend temporally back beyond the individual's existence. The influence of evolution lies of course way back in the past, much beyond the individual's life. Thus the evolutionary process E cannot count as a cognitive process.

[18] The very idea of this criticism can be found in Wielenberg (2014). Wielenberg, however, does not try to re-formulate the argument and, thus, does not arrive at the—very significant—result that will be forthcoming from the following reformulations, namely, that the indirect 'truth tracking' account can be upheld and spelled out by means of a social reactive attitudes story (see section 4, below).

Secondly, premise (2) is unclear. What does it mean to say that 'E is not truth-tracking'? Once we have distinguished between the cognitive process that is responsible for our MAs and the evolutionary process that has significantly influenced our cognitive systems, we have to spell out the 'non-truth-tracking' of the evolutionary process. It is not a cognitive process, and so talk of reliability is not appropriate or must be understood in some other way (different from how it is understood when applied to cognitive systems). How then is the lack of 'truth-tracking' of E to be understood?[19]

Can we save the argument by reformulation?, you may ask. If we want to formulate a *sequitur* argument, we have to reformulate everything in terms of the cognitive process—call it 'C'—which produces the MAs in us. In effect, the recipe for reformulation is quite clear. First, we have to split premise (1) into two, so to speak, one pertaining to the cognitive process C and another one pertaining to the evolutionary process E. Second, we have to find a suitable explication or replacement for premise (2).

3. *Refining the argument*

Following the recipe leads quite naturally to the following reformulation of the argument—call it argument A2:

A2:
(1a) Our MAs are produced by a cognitive process C.
(1b) C has been significantly influenced by E.
(2) E influences cognitive processes significantly because of their fitness contribution.[20]

———

(3) Our MAs are not justified.

The problem with this reformulation, however, is quite clear: it is still a *non sequitur*. The premises are silent about the reliability of the relevant cognitive process C. By merely stating that the evolutionary process E goes by fitness contribution, we have not yet been told whether the result of E's influence is reliable or not. In other words, we have not yet spelled out the first part of the slogan, 'evolution does not care for truth ...', but only the second part, '... but (only) about survival'. Even if we leave in the 'only' it is not clear what follows for the issue of C's reliability. We could change premise (2) accordingly:

[19] A note on the exegetical debate may be in order here. Copp (2008), Enoch (2010), and Skarsaune (2011) struggle with understanding Street's 'tracking account'. But it seems that the issue is undecidable since unclear. Only cognitive processes, or systems, that produce truth-evaluable states are in the business of 'tracking truth', i.e., are supposed to be reliable. The evolutionary process is no such process and, thus, falls outside of any reliable/unreliable classification.

[20] I choose the 'because of' formulation in order to avoid any controversial commitment to teleological notions, such as selection-for. One could equally well speak of 'according to'. Nothing hangs on this.

(2') E influences cognitive processes significantly because of, and only because of, their fitness contribution.

Still it would be unclear what consequences for reliability this would have. Any such consequences should be spelled out explicitly, since they are not obvious at all and are the ones that do the crucial work in the argument. All we are entitled to assume from contemporary evolutionary theory is (1b) and (2) or (2'), but evolutionary theory neither contains nor implies any claim about the unreliability of C.

So how are we to fill in the required additional premise in the best possible way?[21] I submit that the best supplemented argument, argument A3, reads like this:

A3:
(1a) Our MAs are produced by a cognitive process C.
(1b) C has been significantly influenced by E.
(2a) E influences cognitive processes significantly because of their fitness contribution.[22]
(2b) If C has been significantly influenced by E because of a fitness contribution, then C is not reliable.

---

(3) Our MAs are not justified.

Now the problem lies with *premise (2b).* This further premise (2b) is explicitly about reliability or truth-tracking, and so the gap is filled. But (2b) is *not plausible* (as will be shown in a minute). We have turned the argument into a *sequitur*, but only at the price of introducing an implausible assumption. Therefore, the undermining-defeater argument fails. And so the overall conclusion is that no reason against moral realism has been presented so far. Moreover, this result remains even if we switch from (2a) and (2b) to (2a') and (2b'):

(2a') E influences cognitive processes significantly because of, and only because of, their fitness contribution.
(2b') If C has been significantly influenced by E because of, and only because of, its fitness contribution, then C is not reliable.

It remains to be shown that premise (2b) is not plausible. There are at least two reasons. The first reason is that there is a *clear counterexample*, namely, *perception.* Perceptual systems count as cognitive systems in the relevant sense, since they provide perceptual representations which can be correct or incorrect (accurate or inaccurate) and, thus, are evaluable as reliable and unreliable.[23] Quite plausibly, perceptual systems (like the visual system) have been under the significant influence

[21] There might be some unclarity about what it means to say that E exerts influence 'because of fitness contribution', or that E 'goes by fitness'. But whatever unclarity there is, it is a further problem, not the problem I am belaboring here.

[22] (2a) = (2). I have chosen the renaming simply because of the nicer partitioning that results.

[23] I will switch back and forth between processes and systems. Nothing should hang on that.

of evolution, and they have been so because of their fitness contribution. But it is not plausible to think that perceptual systems are *therefore* unreliable. (2b) does not state a true connection between evolution's influence and the reliability of its outcome, the evolved cognitive processes.[24] The second reason for rejecting (2b) is a *theoretical consideration* (that somehow generalizes the first point). In general, a fitness contribution may *consist exactly* in reliably correct representation generated by a cognitive system (of the relevant part of reality, under favorable conditions, and in cooperation with a suitable action-control system, of course).[25] Such reliability might be useful, or even very useful. This applies to our MAs just as well as to any other cognitive systems and processes—as long as we are dealing with truth-evaluable attitudes or states.[26] For example, to have moral attitudes that favor cooperation with cooperative partners is of course very useful for getting along well with other individuals, on the whole and at large. (We will take up and describe this kind of social advantage further in the next section.) Being reliably correct about the moral facts can thus be very beneficial. If there is any truth to this theoretical consideration, then premise (2b) cannot be true. So argument A3 is not sound.[27]

[24] The case of perception as a counterexample has been noticed by others, for example, by Brosnan (2011).

[25] In this connection Peter Graham has put forward very interesting points about the more precise way in which the truth (veridicality, accuracy) of representations can be understood as contributing to the fitness of their possessors, even without those representations having (directly) the (teleo-)function of increasing fitness. See Graham (2014). One of the most important points is that such contributions take the form of a whole package with a functional analysis such that each element does what the other elements need, if everything goes the normal way, in order to produce the fitness increasing behavior. Producing true (veridical, accurate) representations is what the cognitive process C does (in normal conditions), and other elements have the job of producing behavior which is appropriate to the corresponding truths, based on these true representations. This seems to me to be exactly the way in which it could be said that 'evolution *indirectly* cares for truth'. In this way, any suggestion of an incompatibility of adaptation on the one hand and reliably correct representation on the other hand can be rejected. Our cognitive process C can be seen to be both an adaptation (with the function of reliably producing true representations of moral facts) and reliable or 'truth tracking' (in favorable or normal conditions for which it is made). This is exactly *how* C can contribute to fitness.

[26] Teleosemantics even builds an account of representation on this idea of usefulness. There are many versions of teleosemantics that differ in details. But a common core is that the usefulness of (the use of a significant amount of earlier) correct information is partly constitutive of representation (now). Cf., for example, Millikan (1984) and Dretske (1995).

[27] Huemer has tried to present a kind of argument, albeit in a sketchy form (as Huemer himself admits), for the reliability of our moral attitudes or "for why people should have correct ethical beliefs", as he puts it (Huemer 2005: 2). He argues that evolution favors having correct ethical beliefs given that these have a certain, 'altruistic' content (Huemer 2005: 218–9). As I understand his sketchy argument, it in effect amounts to arguing against (2b).

4. *Constraints on moral facts*

Now the situation is the following one. In order to block the argument, the moral realist appeals to the idea that evolution can *indirectly* care for truth (and, plausibly, does indirectly care for truth), namely, in the sense of truth providing a contribution to fitness.[28] We can now ask the question: what do the moral facts have to look like in order for this idea to work? What is required for such a contribution? Does the idea put any interesting constraint on the moral facts? For example, do they have to be 'objective' in any sense? Do they have to be causally efficacious? One could think that there are some interesting necessary and/or sufficient condition that could be derived from the moral realist's idea of reliable representation of moral facts playing the role of a contribution to fitness. Which ones could that be?—Call this 'the constraint question'.

The first, immediate answer to the constraint question that I will argue for is that the moral facts have to be capable of playing a certain role—call it role 'R'. Role R can be described as follows:

(R) The moral facts must be sufficiently reliably representable by the members of the group S and it must be possible that sufficiently many members of S (actively and passively) respect the moral facts.

The first part of (R) is the more important one, for our purposes. It is sufficiently reliable representability of the moral facts. In other words, the moral facts must not be too hard to detect. The members of the group have to have a quite easily available means of reliably representing the moral facts. Of course, they need not be infallible and may make mistakes about the moral facts sometimes. But a sufficient amount of accurate representation must be guaranteed. This the *epistemic condition* that the first part of (R) expresses.

The second, 'practical' part is rather obvious, but it is mentioned since it is required to understand the full story. It concerns the possibility of doing what the moral facts require and of showing the reactive attitudes of praise and blame, and all the rest of the reactive attitudes that have been discussed since Strawson, towards someone who is, or is not, acting in accordance with the moral requirements.[29] If it is a moral fact, for example, that Kim ought to help Jones (because Jones has been injured in some accident), then typically, Kim must be capable of doing what she is required. Call this the 'active respecting of moral facts'. Equally, the members of the group must be capable of showing the positive and negative reactive attitudes towards the agents who act either in accordance or in violation of the moral facts, at least sufficiently often. Call this the 'passive respecting of the moral facts'. Active

[28] For the present purposes, we can count "the four Fs" as what evolution *directly* cares for, as Graham puts it nicely: "feeding, fleeing, fighting, and reproducing" (Graham 2014: 19).

[29] Cf. Strawson (1962).

and passive respecting of moral facts have to be sufficiently possible, says the second, 'practical' part of (R).

Now the new story about how evolution could indirectly favor the reliability of C can be told. It is a *social reactive attitudes story* that fills the gap in the defense of moral realism against the evolutionary debunking attempt. As long as sufficient reliable representation in a group S is secured, and the members of S are sufficiently capable of actively and passively respecting the moral facts, it will in general be of advantage to any individual of S to correctly represent the moral facts and to act in accordance with them. In our example, crudely put, if Kim helps Jones, she will probably be recognized as a kind, supportive person and will receive praise and other positive reactions from the others. The moral attitudes (MAs) tend to favor such helping behavior, and so they indirectly contribute to the positive reactions received from other. (That the members of S are capable of reliably representing the *actions* performed by members of the group must also be secured, of course. They must be able to see whether the action fits the moral facts or not, at least sufficiently often. I take this to be granted since uncontroversial.) It is thus entirely unmysterious how a reliably working cognitive process C governing MAs can contribute to fitness and survival via social interactions and the reactive attitudes therein. Linking positive reactions to conformity with the moral facts can do the job. MAs can benefit their subjects even if they do not benefit because they are true—they can benefit indirectly.[30]

The advantages provided by the social mechanism of reactive attitudes in a group can be quite high, indeed, they can be extremely high. It all depends on how strong the reactive attitudes are. (As we all know, in fact they are quite high nowadays and include all kinds of social exclusion or punishments etc.)[31]

At this point, we can connect the discussion directly to Street's writings. The just-mentioned social reactive attitudes account directly rebuts Street's 'implausible coincidence objection'. According to the moral realist, it is no mere coincidence, unexplicable or mysterious, that our

[30] Note that the proposal is not committed to any substantive normative claim that other third-factor accounts are committed to. For example, Enoch proposes the normative claim that survival and reproductive success are somewhat good (cf. Enoch 2010: 430). Wielenberg assumes the substantive claim that human beings have rights (cf. Wielenberg 2010). Normative claims like these are used by their proponents to argue for the reliability of our MAs. It is not entirely clear, however, how this argument is supposed to run. And, more importantly, it is very controversial whether one can rely on some such morality claim or not, since a quite serious suspicion of circularity or question-begging arises here. (Cf. Vavova 2014, Moon 2016, and Klenk 2017 for discussions on this point).

[31] To connect the discussion directly to Street's writings at this point: The consideration just given directly rebuts Street's 'implausible coincidence objection'. According to the moral realist, it is no mere coincidence, unexplicable or mysterious, that our MAs reliably represent the moral facts, and it is no mere coincidence that their reliably representing these moral facts is beneficial. Cf. Street (2006: 125).

MAs reliably represent the moral facts, and it is no mere coincidence that their reliably representing these moral facts is beneficial.[32]

The first or immediate answer to the question just given is not the end of it. We can go on and ask what further conditions have to be in place in order for truth making the envisioned contribution to fitness. For we can ask what the moral facts have to be like in order to be sufficiently reliably representable by the members of the group. For example, do they have to be 'objective'?

The second answer to the constraint question then is that there is *no further interesting constraint on moral facts* that could be derived from sufficiently reliable representability. As long as the moral facts are sufficiently reliable representable, the social mechanism just described can work. Sufficient reliable representability may have its preconditions. And whatever is required for sufficient reliable representability has to be the case in order for the story of indirect contribution to fitness to work. If we leave to one side the second, practical part of role (R)—which is appropriate in the present context since it is not questioned by any party in the debate—what remains is simply the *epistemic condition of being sufficiently reliably representable*. Succinctly put, as long as the moral facts are *sufficiently epistemically accessible*, the moral realist can make good on the idea that evolution indirectly cares for the truth of moral attitudes.[33]

In order to make this second answer plausible, let us run through a number of candidate conditions that might easily come to mind. Let us begin with *full objectivity*, i.e., total mind-independence. This is not a requirement for the story to work since if the moral facts are like some consequentialist think they are—i.e., an action's maximizing *pleasure* or maximizing *desire* satisfaction, which is belief-independent but not entirely mind-independent—the story is in no way excluded. But fully objective, entirely mind-independent moral facts could fit the same bill. If the moral facts ultimately consisted in some primitive moral reasons relations, holding between some descriptive facts and certain responses (actions and/or attitudes), they could play the very same role—as long as the moral reasons facts are sufficiently epistemically accessible. Next, the moral facts could even be *subjective* in the sense of being *relative* to persons. Suppose that what is morally good or what one morally ought to do varies from person to person. Even this would not undermine the story. As long as the moral facts are sufficiently epistemically accessible (and practicable ...) acting in accordance with the moral facts would be likely to contribute to fitness. Finally, it does also not matter whether the moral facts are *reducible* to descriptive facts or *causally efficacious*. Perhaps, some condition like *supervenience* is

[32] Cf. Street (2006: 125).

[33] It seems appropriate to call this condition 'epistemic' since reliable representation is sufficiently similar to knowledge, and knowability would of course be fine, too.

necessary for sufficiently reliable representation. But causal efficacy does not seem to be required (as the case of knowledge of mathematics shows).[34] And supervenience of the moral on the non-moral facts is a thesis that is widely accepted among moral realists, both naturalistic and non-naturalistic ones. To be sure, reliable representation (or even knowledge) is not for free. Arguably, it requires *some mechanism* which supports the reliable tokening of representations. But the important point is that it is not as though *the social role of moral facts* described in the story required some robust anchoring of moral facts in non-moral, descriptive facts. It is only the *epistemic accessibility* that might require this. And supervenience might be all that is needed for that.

The conclusion of these considerations is thus easily stated. *The one and only constraint* that the role (R) puts on the moral facts is their *epistemic accessibility*. As long as the moral facts are sufficiently reliably representable the social story of reactive attitudes can work and thus 'implement' a way of truth being highly significant to fitness—truth about the moral facts. Therefore, the moral realist has a plausible story about how to explain why premise (2b) in the argument above fails. And the story is fully in line with, and spells out, the idea of evolution caring *indirectly* for the reliability of our MAs. The evolutionary influence on our cognitive process C is fully compatible with its favoring the reliability of C. The MA's *raison d'être* is tied to the reliability of their generating process C. Whether the moral facts have to be objective in any interesting sense can be left open—unless it is entailed by the epistemic accessibility of the moral facts.

If this conclusion is correct, it follows that *the evolutionary debunking argument does not pose any new problem for moral realism*, since the epistemic accessibility of moral facts is an old problem that has long been recognized and discussed. So interestingly, we have been lead back to the old epistemic problem, and no new problem arising from evolution has been discovered.

In sum, the evolutionary debunking argument that has been extracted from the Streetian considerations does not yield a sound argument against moral realism—not even if some objectivity requirement on moral realism is imposed. Once a proper statement of the argument

[34] Here is the right place to critically comment on Mogensen's distinction between proximate and ultimate causes of MAs and corresponding kinds of biological explanation (cf. Mogensen 2015). The application of this distinction matches the distinction between the evolutionary process E and the cognitive process C insofar as the former concerns phylogeny and the latter concerns the individual. But there is no *causal* implication: the cognitive process need not have moral facts as proximate *causes*. So Mogensen's focus is too much on causation, whereas the appropriate focus should be on reliable representation. In addition, the explanatory story on offer here gives moral facts a role in the *phylogenetic, evolutionary genesis* of our MAs—so an 'ultimate' role, if you like—and not (only) in the *individual's* MAs—the 'proximate' role –, as Mogensen wants to have it (cf. Mogensen 2015: 197). Therefore, Severini's criticism of Mogensen's use of the proximate/ultimate distinction does not apply to the story on offer here (cf. Severini 2016).

has been formulated—argument A3—the crucial weakness becomes apparent: evolution's pressure towards fitness does by no means exclude that reliable cognitive processes will be favored. Quite the contrary, it seems that evolution *indirectly* cares for truth, or at least can indirectly care for truth. The moral realist can tell a plausible explanatory story about how this influence could have developed, the story of reactive attitudes. Thus, the response to the debunking argument is no longer just an in-principle possibility of 'indirect truth caring' but a concrete, though sketchy, explanatory account of how this in-principle possibility can be realized. In addition, we can see what the moral facts have to be like in order to play their role in this story. They simply have to be sufficiently epistemically accessible, and no more. The evolutionary considerations, therefore, bring out the importance of the epistemology of moral facts. Some solution to this epistemological challenge has to be found. But this is not a new problem, and so the evolutionary argument does not yield any new constraint on moral realism.[35]

References

Alston, W. 1993. *The Reliability of Sense Perception*. Ithaca: Cornell University Press.

Alston, W. 1986. "Epistemic circularity." *Philosophy and Phenomenological Research* 47 (1) 1–30.

Bergmann, M. 2006. *Justification Without Awareness*. Oxford: Oxford University Press.

Brosnan, K. 2011. "Do the evolutionary origins of our moral beliefs undermine moral knowledge?" *Biology and Philosophy* 26 (1): 51–64.

Burge, T. 2010. *Origins of Objectivity*. Oxford: Oxford University Press.

Copp, D. 2008. "Darwinian skepticism about moral realism." *Philosophical Issues* 18: 186–206.

Crow, D. 2015. "A Plantingian pickle for a Darwinian dilemma: evolutionary arguments against atheism and normative realism." *Ratio* 24 (2): 130–48.

Cuneo, T and Shafer-Landau, R. 2014. "The moral fixed points: New directions for moral nonnaturalism." *Philosophical Studies* 171: 399–443.

FitzPatrick, W. J. 2015. "Debunking evolutionary debunking of ethical realism." *Philosophical Studies* 172: 883–904.

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

Enoch, D. 2010. "The epistemological challenge to metanormative realism: how best to understand it, and how to cope with it." *Philosophical Studies* 148: 413–38.

Goldberg, S. 2010. *Relying on Others*. Oxford: Oxford University Press.

[35] I am very grateful for helpful comments and discussions to Christoph Fehige, Thomas Grundmann, Susanne Mantel, and Peter Schulte. An earlier version of the paper has been presented in the Colloquium in Practical Philosophy at the Saarland University, Saarbruecken, Germany, organized by Christoph Fehige and Ulla Wessels, in January 2015. Many thanks to the organizers and participants for stimulating comments and valuable points of criticism.

Goldman, A. 1979. "What is justified belief?" G. S. Pappas (ed.). *Justification and Knowledge*. Dordrecht: Reidel: 1–23.

Goldman, A. 2008. "Reliabilism." *Stanford Encyclopedia of Philosophy*, ed. Ed Zalta.

Graber, A. 2012. "Medusa's gaze reflected: a Darwinian Dilemma for anti-realist theories of value." *Ethical Theory and Moral Practice* 15: 589–601.

Graham, P. 2014. "The function of perception." In A. Fairweather (ed.). *Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science*. Synthese Library 366: 13–31.

Huemer, M. 2005. *Ethical Intuitionism*. London: Palgrave Macmillan.

Hofmann, F. 2013. "Three kinds of reliabilism." *Philosophical Explorations* 16 (1): 59–80.

Joyce, R. 2007. *The Evolution of Morality*. Cambridge: Cambridge University Press.

Kahane, G. 2011. "Evolutionary debunking arguments." *Nous* 45 (1): 103–25.

Kelley, T. 2014. "Evidence." *Stanford Encyclopedia of Philosophy*.

Klenk, M. 2017. "Can moral realists deflect defeat due to evolutionary explanations of morality?" *Pacific Philosophical Quarterly* 98: 227–48.

Kyriacou, C. 2016. "Are evolutionary debunking arguments self-debunking?" *Philosophia* 44: 1351–1366.

Pollock, J. 1986. *Contemporary Theories of Knowledge*. Lanham: Rowman and Littlefield.

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories*. Cambridge: MIT Press.

Mogensen, A. 2015. "Evolutionary debunking arguments and the proximate/ultimate distinction." *Analysis* 75 (2): 196–203.

Moon, A. 2016. "Debunking morality: Lessons from the EAAN Literature." *Pacific Philosophical Quarterly* 98: 208–26.

Ruse, M. and Wilson, E. O. 1995. "Moral philosophy as applied science." In E. Sober (ed.). *Conceptual Issues in Evolutionary Biology*. Cambridge: MIT Press: 421–38.

Senor, T. S. 1996. "The prima/ultima facie justification distinction in epistemology." *Philosophy and Phenomenological Research* 56: 551–66.

Severini, E. 2016. "Evolutionary debunking arguments and the moral niche." *Philosophia* 44 (3): 865–75.

Skarsaune, K. O. 2011. "Darwin and moral realism: survival of the iffiest." *Philosophical Studies* 152: 229–43.

Strawson, P. 1962. "Freedom and resentment." *Proceedings of the British Academy* 48: 1–25.

Street, S. 2008. "Reply to Copp: naturalism, normativity, and the varieties of realism worth worrying about." *Philosophical Issues* 18: 207–28.

Street, S. 2006. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127: 109–166.

Smith, M. 1994. *The Moral Problem*. Oxford: Oxford University Press.

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

Stroud, B. 1984. "The problem of the external world." In B. Stroud. *The Significance of Philosophical Skepticism*. Oxford: Oxford University Press: 1–38.

Vavova, E. 2015. “Evolutionary debunking of moral realism.” *Philosophy Compass* 10 (2): 104–16.

Vavova, E. 2014. “Debunking evolutionary debunking.” *Oxford Studies in Metaethics* 9: 76–101.

Wielenberg, E. J. 2010. “On the evolutionary debunking of morality.” *Ethics* 12 (3): 441–64.

Wielenberg, E. J. 2014. *Robust Ethics*. Oxford: Oxford University Press.

Croatian Journal of Philosophy
Vol. XVIII, No. 54, 2018

On a Consequence in a Broad Sense

DANILO ŠUSTER
University of Maribor, Maribor, Slovenia

Cogency is the central normative concept of informal logic. But it is a loose evaluative concept and I argue that a generic notion covering all of the qualities of a well-reasoned argument is the most plausible conception. It is best captured by the standard RSA criterion: in a good argument acceptable (A) and relevant (R) premises provide sufficient (S) grounds for the conclusion. Logical qualities in a broad sense are affected by the epistemic qualities of the premises and "consequence" in a broad sense exhibits an interplay of form and content. There are four proposals for the premise—conclusion relation: (i) no strictly logical connection ("non-logical" consequence); (ii) one type of connection only (deductivism); (iii) a few types of connection (deduction, induction, perhaps conduction and analogical reasoning); (iv) many types of connection (argumentation schemes). Deductivism is a serious option but in its strong version, as the discussion about petitio shows, it fails to establish that arguments which are not cogent are thereby invalid. And weak deductivism, very attractive from the pedagogical point of view, has some deficiencies (implausible hidden premises; preservation of truth, not probability). I argue that the idea of a counterexample, when we regard certain components of the argument as fixed and others as variable, is the best approach to the analysis of the illative core of every-day arguments (the approach of David Hitchcock on material consequence).

Keywords: Informal logic, consequence, begging the question, cogency, deductivism, counterexample.

1.

W. V. O. Quine (1950: vii) opens his *Methods of Logic* with a famous quote: "Logic is an old subject, and since 1879 it has been a great one."[1*] The year marks the appearance of Frege's *Begriffsschrift* and the im-

[1*] The writing of this paper was supported by the *Croatian Science Foundation* under the grand number IP-06-2016-2408 and the *Slovenian Research Agency* research core funding No. P6-0144.

pressive development of modern logic ever since. But there is *Logic* (the old subject) and there is the *logical* dimension of (everyday) argumentation and reasoning. Often about hot political issues. Thus in his notorious *Diary* Frege (1924) writes about patriotism:

> The question here is not about a judgment in the sense of logic, not about considering something as true, but about one's feelings and inner attitude. Only Feeling [Gemüt] participates, not Reason, and it speaks freely, without having spoken to Reason beforehand for counsel. And yet, at times, it appears that such a participation of Feeling is needed to be able to make sound, rational judgments in political masters. (Mendelsohn 1996: 33)

The following comment is perhaps too harsh: "The man who wanted to set mathematics on surer logical foundations, was content for politics to be based on emotional spasms." (Monk 2017). Still, these are surprising claims for the founder of modern logic which make one wonder how do formal logical theories and the logic of every-day reasoning mesh together. The latter is nowadays the subject of the so-called "informal" logic, characterized rather broadly as a "collection of normative approaches to the study of reasoning in ordinary language that remain closer to the practice of argumentation than formal logic" (van Eemeren 2009: 117). Originally the opposition to formal logic was more clearly stated:

> … that branch of logic whose task it is to develop non-formal [i.e., not restricted to logical form] standards, criteria, procedures for the analysis, interpretation, evaluation, critique and construction of argumentation in everyday language. (Blair 2014: 373–374.

One of the pioneers of the informal logic later adds (Blair 2015: 27): "I would today drop 'standards,' and say "arguments and argumentation" and "natural language"". I agree with the ecumenical spirit of the remark—classical deductive standards are no longer excluded by fiat (the original definition was: "Informal logic designates that branch of logic whose task is to develop non-formal standards, criteria, procedures for the analysis, interpretation, evaluation, critique and construction of argumentation in everyday language" (Johnson and Blair 1977: 148). But the working assumption still seems to be that the analysis of arguments and argumentation in natural language has little to do with the areas of formal logic where Frege made his great contributions.

There is another issue where Frege's approach was described as having "deleterious effects both in logic and philosophy" (Dummett 1973: 432–433). According to Frege in logic truth is not merely the goal, but also the object of study. Traditionally, however, the relation of logical consequence ("transitions from sentences to sentences") is the proper subject-matter of logic. "Informal" logicians speak about the premise-conclusion relationship as the "illative" core of argumentation, "This, therefore that," a single integrated set of one or more propositions adduced as grounding or evidence in support of a claim. An illative move or a series of illative moves is made "… from the basis or starting point of the reasoning or argument to the upshot that is inferred or alleged to follow from that basis. Some call this move an inference, others call it an implication, others call it a premise-conclusion link, and others call it a consequence relation" (Blair 2012: 103).

How to characterize this "illative" core of argumentation from the "informal" or broad point of view?

2.

In contrast to classical *soundness*, requiring valid arguments with true premises, *cogency* emerged as the central normative notion in the approaches that remain closer to the practice of argumentation. Unfortunately the notion is not well defined and the usage is not uniform. Some use it *broadly* to cover the qualities of a successful argument, others use it *narrowly* as a characterization of good reasoning (strictly illative moves). Moreover, there are subdivisions within each camp, narrow usage encompasses either inductive strength (corresponding to deductive validity) or both deductive and non-deductive patterns of reasoning. In the other camp some will reserve the label for something like inductive soundness (corresponding to classical soundness), while others speak of *all* of the qualities of a successful argument (deductive or non-deductive). Consider the scheme:

COGENCY	**Narrow**	**Broad**
Reasoning	Inductive strength	Inductive and deductive Umbrella validity (Govier)
Argument	Inductive "soundness"	Good

The narrowest option (inductive strength corresponding to classical validity) is exemplified, for instance, by Feldman (2014: 95): "an argument is cogent if and only if it is not valid but the premises of the argument are good reasons for the conclusion," or "an argument is cogent if and only if it is not valid but the conclusion is probably true if all the premises are true." Broader, but still limited to non-deductive arguments are typical uses in contemporary critical thinking literature, for instance: "a cogent argument is an inductive argument that is strong and has all true premises" (Hurley 2015: 52 and Baronett 2015: 43). Cogency is "inductive soundness" so to speak. But the textbook usage is not uniform at all, sometimes an idiosyncratic terminology is used: "An argument is *reliable* when it is inductively strong and has all true premises" (Johnson 2016: 10). Or Vorobej (2006: 54): "An argument is reliable just in case both (a) it is not valid and (b) its conclusion is more likely to be true than false, given that each of its premises is true." The classic Copi textbooks do not define the notion at all, though they do speak about arguments being *fairly* cogent or *moderately* cogent (Copi 1990: 538), qualifications which make sense for inductive strength only.

Cogency as an illative evaluation of the reasoning in a *broad* sense is supposed to cover both arguments which are deductively valid and those which are inductively strong. According to Cozzo (2017): "A cogent inference is an inference that "compels us to accept the conclusion" if we accept the premises." Govier (2018: 288) introduced the notion of *umbrella* validity: "An argument is (umbrella) valid if its premises are properly connected to its conclusion and provide adequate reasons for it." Plumer (2016: 92) stipulates that cogency should be used instead: "I take it to pertain only to an argument's reasoning or logic, not also to the truth value of its propositional elements (unlike the technical concept of soundness) … I take cogency to be the broader notion of proper reasoning as compared to the technical concept of validity." But then, somewhat surprisingly, he adds (Plumer 2016: 92): "Depending on how the constituent notions are explicated, we can agree with Johnson & Blair's (1977) well-known and widely accepted "RSA" criteria for argument cogency: the premises are to be relevant, sufficient, and *acceptable*."

The relevance of premises and their sufficiency pertain to the adequacy of the (broad) inferential link: the reasons offered must be probatively relevant to the conclusion and they have to be sufficient for accepting it. The relevance "criterion" is best understood as a criterion of *inclusion* of premises in the analysis and reconstruction of arguments. Only probatively relevant propositions may be counted as premises (Blair 2012: 93). One might say that relevance establishes the inferential *connection* and sufficiency guarantees its *strength*, so you can have relevance without sufficiency (weak connections—typical hasty generalizations). But the converse is not possible, if the premises are not relevant, they cannot be sufficient either (cf. Biro and Siegel 1992). Although there is still some discussion about whether to require truth or acceptability as a condition of premise adequacy (Johnson 2000: 195–199), I agree with the mainstream which favours acceptability over truth. Real life arguing often takes place in contexts characterized by uncertainty (hypothetical and uncertain beliefs, deep disagreements about what is true and false, ethical and aesthetic claims) where the truth is too stringent (or inappropriate). Also, there had better be a sense in which false conclusions can sometimes be reasonably well supported—the whole discussion about pessimistic induction in philosophy of science (the cases of generally accepted but false theories in the past which are supposed to subvert our expectations about our best present-day theories) would otherwise be pointless. Acceptability is an epistemic notion, roughly, premises are acceptable when it is reasonable for those to whom the argument is addressed to believe them (they are justified in believing them).

A negative designation of cogency based on the RSA criteria is now also an option—to be cogent the argument must avoid three basic fallacies: irrelevant reason, hasty conclusion and problematic premise (cf. Freeman 2011: xi). The criterion now amounts to the broad notion of soundness and we thus get cogency in the broadest possible sense. Here are some variations. According to Adler (2006: 225): "'Cogency' is used broadly to refer both to correct support relations, validity, in the case of deductive arguments, and to the soundness, warrant, and relevance of the premises." But he then adds: "I use 'cogency' as a generic term to cover the qualities of a successful argument." I suppose this is the dominant view in the theory (if not in the practice exemplified by the textbooks) of informal logic, witness Govier (2018: 287–88): "If the premises of an argument are rationally acceptable and are ordered so as to provide rational support for the conclusion, the argument is cogent." Or Blair (2012: 46), "A logically 'cogent' argument has acceptable premises as well as an acceptable premise-conclusion link" and Hitchcock (2017: 4): "I take an argument to be cogent for somebody when and only when (1) that person has justifications which are independent of the conclusion for accepting its premises and (2) the conclusion follows from the premises." This is also close to typical philosophical usage, thus Wright (2002: 331): "cogent argument is one whereby someone could be moved to rational conviction of-or the rational overcoming of doubt about-the truth of its conclusion."

The last option, cogency as a generic term covering all of the qualities of a good, well-reasoned argument seems to me to be the best choice. Narrow (inductive) reasoning is just too restrictive. Why should a simple *modus ponens*, for instance, when considering whether to water the garden, my wife says: "It's going to rain. If it is going to rain, there is no need to water the garden," not be a cogent everyday argument? A deduction which never the less moves me to the rational overcoming of doubt about the truth of its (omitted) conclusion. The same considerations will exclude inductive soundness as too narrow. And cogency as broad reasoning implies that valid arguments are *always* cogent, but the notorious question-begging arguments are valid, yet they lack the qualities of being good and cogent arguments, as we shall later see.

Still, we should acknowledge the fact that cogency is a loose evaluative concept and its meaning, as the examples above show, is to some degree stipulative (Plumer 2016: 92). But I think that the oscillation between good *reasoning* (strictly illative moves only) and good *argument*, one which deserves to convince us of its conclusion, marks one of the central turning points of informal logic with respect to classical formal logic. Will an argument fail to be cogent if you have no reason to believe one or more of its premises? Can you have a *well*-reasoned argument with unacceptable premises? Can evidential considerations affect the quality of reasoning in the broad sense? The majority of informal logicians would say *yes*. They frequently view the determination of the *acceptability* of premises as an important part of the logical appraisal of arguments. Whether the relevant premises warrant a conclusion depends on what else is known about the matter under consideration. Plumer (2006: 93) quotes Salmon that nondeductive reasoning is cogent if "the argument has a correct form, and … the premises of the argument embody all available relevant evidence." And according to Pinto (2001: 27): "assessment of inferential link cannot be carried on in isolation from assessment of premise acceptability." I agree, just consider some typical instances of absent evidence reasoning (or arguments from ignorance):

> I checked the train table: the connection between Ljubljana and Venice is not listed. No records, so no train connections? Marco Polo's travel journals are silent on the Great Wall of China. No evidence, so no visit? If evolution happened, where have all the intermediate forms gone? No fossil records, so no evolution? The fact that no one has been able to pick up a tailpipe from a UFO does not mean UFOs do not exist. Absence of evidence is not evidence of absence? A wave of recovered memories about alien abductions is likely the product of fabrication or suggestive therapeutic techniques, because we have never found any material traces of these alien abductions. No evidence, so no abductions? There is no reliable evidence available to us of the number of stars being even. So it must be odd.

Some of these pieces of reasoning are cogent, some are fallacious, but they display a typical interplay of form and content, logical qualities in a broad sense are affected by epistemic qualities of premises.

3.

I began with the normative issue—the logical evaluation of the "illative" core, but so far I said nothing about the *nature* of this core. Predictably, there is no consent, the informal community is working with the following proposals for the premise—conclusion relation: (i) no strictly *logical* connection ("non-logical" consequence); (ii) one type of connection only (deductivism); (iii) a few types of connection (deduction, induction, perhaps conduction and analogical reasoning); (iv) many types of connection (argumentation schemes).

As for the first option, one could start with classical deniers, say Quine (1986, vii) and his dismissal of the application of the word 'logic' as covering both, deductive and inductive logic: "The philosophy of inductive logic, however, would be in no way distinguishable from philosophy's main stem, the theory of knowledge". *Informal* logic as a separate approach was more visible for Hintikka (1999: 115): "I have a great deal of sympathy with the intentions of those philosophers who speak of 'informal logic', but I don't think that any clarity is gained by using the term 'logic' for what they are doing." One can still find claims that "Nonformal logic is the science of arguments not strictly governed by consequence" (Hanna 2006: 30) and even informal logicians themselves, in the spirit of rejecting formal logic as the tool to be used for the analysis of natural language argumentation, sometimes characterize illative moves as "non-logical" consequence (Hitchcock 2009). These claims are based on a certain narrow conception of logical form and logical consequence—what is distinctively logical about arguments is associated to their *formal* aspects, where individual arguments are valid only in virtue of instantiating truth-preserving logical forms investigated by formal deductive logic. But why exclude the clearly *logical* dimension of everyday argumentation from the domain of logic in a broad sense? To be fair, Hitchcock has done a lot to clarify the general notion of "follows from" but he later calls it, more aptly, *material consequence*.

On the other extreme one finds a growing collection of argumentation schemes—"forms of argument (structures of inference) that represent structures of common types of arguments used in everyday discourse, as well as in special contexts like those of legal argumentation and scientific argumentation" (Walton 1996: 1). Examples are means-end reasoning, inference to the best explanation, inductive generalization from instances, reasoning from the results of a randomized trial to a causal conclusion, lack-of-knowledge arguments, and so forth. Walton initially discussed 25 schemes, but Walton, Reed and Macagno (2008) later identify 96 distinct argumentation schemes. A lot of important work has been done within this approach, but the inflation of presumptively good patterns schemes (like the inflation of "bad" patterns within the fallacy approach to informal logic) has not really helped to clarify the nature of the "following from" relation. Especially since we can, apparently, multiply schemes indefinitely. Do all of these patterns share a common logical core or not?

Two options remain: deductivism and the approach, nowadays dominant, which recognizes various degrees or kinds of the premise-conclusion connection. Govier (1992: 393) calls the last approach the *pluralist* view of cogency, though it is clear that pluralism encompasses a very *limited* number of relations: deductive entailment, conducive support, inductive support and analogy.

Deductivism is the view that ordinary arguments are best analysed as deductive inferences, but this does not mean that the analysis and appraisal of arguments is based upon classical *logical* form and this or that formal system. All defenders of deductivism agree that an inference is deductively valid if and only if it is impossible for its premises to be true and its conclusion false. But they add that not every aspect of good reasoning boils down exclusively to classical soundness. According to the *weak* version deductive validity should not be equated with formal validity: *material* validity will do just as well. Given the premises: "Ann is taller than Bill and Bill is taller than Mary" it is impossible that it should be false that "Ann is taller than Mary." This impossibility is explained in terms of the meanings of non-logical terms ("being taller than") not in terms of standard logical constants. According to *strong* deductivism, however, a principal factor in distinguishing good from bad reasoning is *inferential* deductive validity where an inference is deductively valid if and only if it is *logically* impossible for its premises to be true and its conclusion false.

Jacquette (2007 and 2009) defends strong deductivism, all and only good reasoning is, minimally, deductively valid inference:

> According to deductivism, formal logic is therefore the continuation of informal logic by more rigorous symbolic mathematical methods, while informal logic is the continuation of formal logic by non-symbolic nonmathematical means (2009: 189). /…./ There is but one logic, then, whose gold standard is deductive validity, with purely formal and purely informal logical methods appearing at the extremes of a spectrum of ways of understanding the deductive validity status of inference (2009: 192).

But he immediately faces the problem that valid arguments with true premises are always sound, but not always cogent. There are seem to be instances of fallacious reasoning which are deductively valid and Jacquette readily accepts the challenge: "A single deductively valid informal fallacy is sufficient as a fatal counterexample to deductivism" (2009: 190). He never the less tries to defend deductivism by treating all recognized informal fallacies as *deductively* invalid. A discovery of a single deductively valid informal fallacy or of a cogent but deductively invalid reasoning would present a counterexample to strong deductivism. I will critically discuss the first option only. And one fallacy only—I think that circular reasoning or *petitio principii* is a touchstone for strong deductivism.

Jacquette attempts to *reconstruct* begging the question as a deductively invalid piece of reasoning. The full content of circular reasoning for him is not: "P, therefore P" but rather "P, therefore it is significant (worthwhile, informative) to conclude that P" (Jacquette 2009: 203–204). According to this expanded reconstruction it is logically possible for the assumption to be true and the conclusion false—uninformative and insiginificant (Jacquette 2009: 204): "the thinker falsely supposes that it is significant, worthwhile or informative to conclude that a certain proposition is true from an assumption base that includes the very same true or false proposition." Jacquette acknowledges the fact that it may be an *informal* matter to judge the relevance of the conclusions in question.

According to this diagnosis it is always possible that in circular arguments the premises are true but it is still false that it is significant, worthwhile or informative that the conclusion is true. Of course, *every* traditional fallacy of relevance will automatically fit this bill of invalidity (appeal to force, ad hominem, straw man, missing the point, red herring ...)! This looks like a very cheap victory for deductivism and almost trivial. Defenders of deductivism can be more informative. *Weak* deductivism claims that "natural language arguments should be understood as attempts to formulate deductive arguments" (Groarke 1999: 2). This claim is perfectly compatible with the RSA criterion of cogency. The difference between cogent and fallacious arguments is then to be found in the truth or plausibility of their premises. *Petitio*, though valid, is not cogent because the premises are not *acceptable* (for the audience in doubt of the conclusion).

Jacquette has to offer a different diagnosis: *petitio* is not valid because the conclusion is not inferred *significantly*. Now, being "significant" is on a different level than being "true" (or acceptable), what we have is a normative assessment of reasoning and as we remember from Carroll and his Tortoise (1895), it is never a good policy to mix the levels. First of all the explicit form cannot be just "P, therefore it is significant to conclude that P." It might be the case that for a certain type of audience it is significant to conclude that P and resolve a certain issue, perhaps to justify P by some other reasons. The intended reading must then be: "P, therefore it is significant to conclude that P from P." In order to assess the validity of *this* reasoning we now need a criterion of significance. The conclusion ("It is significant to conclude that …") will then be false either because it violates certain dialectical (rhetorical) norms or because it violates norms of cogency. In any case deductive *invalidity* is not doing any work at all—one could just as well drop the initial *P* from "P, therefore it is significant to conclude that P from P," and explain why it is not significant (informative, etc.) to conclude that P from P! But the explanation will not appeal to the notion of deductive validity.

Moreover, strong deductivism is in danger of falling into the old trap of proclaiming *all* deductively valid arguments as question-begging. All that Jacquette (2009: 204) has to say about this old conundrum is: "The same lack of significance need not plague logically more complex deductively valid inferences, such as *modus ponendo ponens* or *tollendo tollens*, *reductio ad absurdum*, or the like, if these inferences are considered as issuing in worthwhile or informative conclusions." Well, what is the difference? To infer, say, "P & Q, therefore P" is presumably not significant. But "P & Q" is logically *equivalent* to "Q & (Q => P)", so why should *modus ponens* "Q, Q => P, therefore P" be any better in terms of significance? The selection of premises obviously plays an important role. But why so?

Jackson (1987) makes an interesting proposal. By propounding an argument I offer to my audience not only premises as evidence for the conclusion but, in an implicit way, also reasons (evidence) for the *acceptability* of those premises. To take his example:

A Mary is at the party. If Mary is at the party, Fred is too. So, Fred is at the party.

The hearer is entitled to infer that I have *separate* evidence for each of the premises. Perhaps I have just seen Mary at the party, and I also know that Mary and Fred always go to parties together. The way of presenting my argument and the selection of premises provide important information about the evidence available for possible "borrowing." The hearer knows enough about the kind of evidence likely to lie behind my assertions (perception, familiarity with the couple) to borrow it to good purpose. Now take:

B Mary and Fred are at the party. So, Fred is at the party.

In general, to infer "P & Q, therefore P" is not significant, or, as Jackson would say only "marks time." But this need not be true for "Q, Q => P, therefore P" (the form of our first argument). The difference will be explained in terms of the kind of (implicit) reasons I have for the *acceptability* of my premises and the hearer of the argument then borrows. If she doubts the conclusion in the second case (B) she will very likely have background beliefs relative to which the reasons indicated by propounding the argument (seeing them both at the party) will have no impact. This is precisely Jackson's definition of *begging the question* (1987, 35): "an argument such that any (sane) audience which was in doubt about the conclusion would have background beliefs relative to which the evidence provided by propounding the argument has no impact." Note: reasons (evidence) for *accepting* the premise are decisive for the question whether the argument is fallacious or not. Evidential considerations affect the quality of reasoning in the broad sense.

4.

Jackson emphasizes a dialectical and pragmatic dimension of propounding an argument—the *persuasive* power of the argument depends on the impact of the evidence implicitly offered for borrowing on the particular *audience*. This might lead to a different diagnosis of *petitio* and perhaps another escape route for deductivism. Circular reasoning "is not fallacious in the true sense of the word, but objectionable and to be avoided in argumentation for another reason" (Jacquette 2009: 203). *Petitio principii* is generally lacking in argumentative significance, but this alone does not make it fallacious, this form of reasoning remains *valid*. This strategy is in line with contemporary rhetorical and pragma-dialectical approaches to argumentation. Thus Perelman (Perelman and Olbrechts-Tyteca 1971: 112): "the *petitio principii*, which does not concern the truth but the adherence of the interlocutors to the presupposed premises, is not an error of logic, but of rhetoric / … / an error in argumentation."

Crudely put—an *argument* is a set of statements or propositions or natural-language declarative sentences one of which is the conclusion, the remainder of which are the premises. *Argumentation* is the activity of arguing, a complex, social speech act in which either only one speaker presents a thesis to an audience and defends it or more speakers do so "dialectically." According to epistemic theories the principal goal of argumentation is, roughly, to induce belief or elicit a reasoned change in view (Harman 1986). Perelman defends a different, *rhetorical* theory of argumentation—the goal of argumentation is to cause or increase the *addressee's* belief in the conclusion. And *consensus* theories of argumentation see argumentation as a means for reaching consensus, or, in a more elaborate way (Eemeren and Grootendorst 2004: 1):

> Argumentation is a verbal, social, and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of propositions justifying or refuting the proposition expressed in the standpoint.

Circular arguments, in general, are fallacious because they violate normative rules of dialogue which demand consensual starting points. Fallacies are bad arguments in the sense of being Gricean failures of co-operation which violate rules of critical discussion. There are eight such rules and the sixth rule (the starting point rule) states (Eemeren and Grootendorst 2004: 193):

> Discussants may not falsely present something as an accepted starting point or falsely deny that something is an accepted starting point.

By falsely presenting something as a common starting point, the protagonist tries to evade the burden of proof. The techniques used for this purpose include advancing argumentation that amounts to the same thing as the standpoint. Consider *The Bank Manager Example*, "a staple of many textbooks":

> Manager: Can you give me a credit reference?
> Smith: My friend Jones will vouch for me.
> Manager: How do we know he can be trusted?
> Smith: Oh, I assure you he can.

In this dialogue one person is supposed to vouch for the reliability of the other. The reliability of the vouchee is in doubt and some secure source is needed to reassure this doubt. But if the reliability of the voucher is questioned, the reliability of the vouchee cannot be used to reassure this doubt, because it is itself in doubt, in the first place (cf. Walton 1991: 248). One could as well say that Smith falsely presents his reliability as an accepted starting point in a dialogue. But now consider the famous Moore's argument for the existence of an external world (Jackson 1987: 35):

> M1: This is a hand.
> M2: A hand is an external object.
> Therefore: At least one external object exists.

According to the pragma-dialectical approach dogmatist (Moore) *falsely* presents his hands as an accepted starting point (as an object in the external world) in his dialogue with the skeptic.

I cannot discuss all of the nuances of this approach, let us just ask ourselves, *why* is the first premise a false move in the Moore's case? And what differentiates *petitio* from other unacceptable starting points (say inconsistent, irrelevant or doubtful premises)? The discussant who in the discussion fulfils the role of protagonist of a standpoint will in the argumentation stage at a certain moment express a proposition that he claims can be identified as a common starting point by means of the "intersubjective identification procedure." But how will this procedure look like? When the premise is not equivalent to the conclusion it is not at all easy to identify common starting points. I think that the falsity will be revealed through reasoning in the *broad* sense. To continue in line with Jackson—what matters is not just the premises themselves, but the reasons offered for their *acceptability*: M1 is supported by perceptual experience. The sceptic, doubtful about the conclusion, will point out that it is seriously possible that there are no external objects, since we are, say, envatted and handless brain-in-a-vats, having non-veridical sensory experiences. This background will *block* the perceptual reasons for M1 and thus make this premise ineffective. In any case the diagnosis of the falsity is *epistemic*: the premise in the examined case of arguing is epistemologically unsuitable for the purpose of proving (justifying) the conclusion in that particular discussion. And, therefore, we may add, an unacceptable starting point.

Pragma-dialectical approach is perhaps inspired by Aristotle—in the *Topics* he is concerned with contentious disputation between two or more parties. *Begging the question* is said to occur where a questioner, the party who is supposed to be arguing for a certain thesis, asks to be granted the thesis as a premise to be conceded by his opponent. Aristotle uses the same terminology in the *Prior Analytics* (64b 33), where he says it is the attempt to prove what is not self-evident by means of itself. But demonstration proceeds from what is more certain or better known: if a man tries to prove what is not self-evident by means of itself, he begs the original question (64b 37). To beg the question is to violate the *epistemic* principle of the priority in knowledge of the premises over the conclusion in a demonstration. This second account is epistemic, the first dialectical or conversational. Sosa (2004: 57) suggests to use "vicious circularity" in the first case and reserve "begging the question" for something involving not so much proper reasoning as proper dialogue. But it is clear that Jacquette and pragma-dialecticians aspire for a *uniform* explanation of all of the cases in terms of violating certain pragmatic rules. Unsuccessfully, as I have tried to show.

There are various other ways of how to disqualify question-begging arguments as not cogent. According to Woods (2004: 34) "p, so p" is always a fallacious inference but there is nothing wrong with the *entailment* "p entails p." Plumer (2016: 92) declares such arguments as cogent and fallacious (well- and poorly reasoned) at the same time in different respects. But if question-begging arguments are not cogent because the inferential link is defective then cogency incorporates epistemic considerations. In the simplistic formulations above some premise of the argument is *equivalent* to the conclusion. I believe that the *dependency* conception, illustrated by Moore's proof, is more general (cf. Walton 2006). Normally the "flow of inference" in an argument is from the premise to the conclusion. But where it is also required that an inference be made in the other direction, from the conclusion to the premise, the argument begs the question. In every argument the conclusion depends, justificatorily, on the premise, but when the "flow of justification" goes in both directions, the argument begs the question. *Blockades* are also part of the "fallacious" inferential game: doubts about the conclusion might prevent the premise of having any *inferential* power. In any case *petitio* violates the normative requirements of good *reasoning* in a broad sense, it is "fallacious in the true sense of the word," not just pragmatically inappropriate.

5.

If strong deductivism is true, then reconstructions of the informal fallacies (violations of the RSA criteria) as deductive invalidities are possible in every case. I argued that *petitio* remains "a fatal counterexample to deductivism." Weak deductivism, however, remains a viable option for the premise—conclusion relation. Remember: deductive validity is not defined by "formal validity" as canonized in a certain formal system. An argument is deductively valid if (and only if) it is impossible for the premises to be true and the conclusion false, and *material* validity will do as well. Deductivism within informal logic also "recognizes that the domain of premise/conclusion relations is only one ingredient of good argument, and that it is an ingredient which needs to be situated in a more comprehensive account of argument which includes an account of differences of opinion, standpoints, implicit and indirect argument components, and so on" (Groarke 1999: 5). Pragma-dialecticans actually *embrace* deductivism in the form of indirect speech acts expressing hidden premises which make arguments valid.

How to situate weak deductivism with respect to cogency? We might say with Govier (1992: 393) that an argument is cogent if and only if (1) its premises (explicit and implicit) are acceptable to the audience to whom the argument is addressed; (2) its explicit premises, when properly supplemented by implicit premises, deductively entail its conclusion. When the premises of an argument deductively entail its conclusion, that argument satisfies the relevance and sufficiency conditions according to Govier (2010: 90). This is slightly imprecise—I agree with Hitchcock that a deductive argument still establishes its conclusion if it contains an irrelevant premise; it is simply inelegant because of this superfluity (Hitchcock 2017: 361). Still, let us assume that deductive arguments are unobjectionable from the 'R' and 'S' point of view. But, as we saw, cogency includes acceptability and for Govier (2018: 430) at least, question-begging arguments "will be adequate from the point of view of deductive logic, and yet be *inferentially* flawed because the audience cannot rationally move from acceptance of the premises to acceptance of the conclusion."

Weak deductivism is a very simple theory—the inference relation is an all-or-nothing thing. For the opponent, to use a metaphor suggested by Groarke (2009: 102), the inference relation is like glue which comes in different strengths: "Sometimes premises and conclusion are glued so tightly together, the bond is almost unbreakable; sometimes the bond is extremely weak and tenuous; sometimes, somewhere in-between." I think that the strongest case for deductivism comes from pedagogical practice. There is only one type of reasoning and instructions for the reconstruction of natural language arguments are very simple:

look for additional premises that explicitly link the original premises to the conclusion in such a way that the reconstructed argument comes out as valid. The whole burden of evaluation is then on the acceptability of the premises. This comes as a relief for anybody engaged in teaching informal logic and critical thinking where one often wonders what kind of techniques, exactly, to teach and how to test the results.

Attractive as it is deductivism also has some well-known deficiencies. Many arguments appear to offer reasonable, but not deductively conclusive justification for their conclusions, yet the hidden premises needed to make them valid are just too strong and so unacceptable. Consider the very mundane case discussed by Groarke (2009: 97): "The weather network said it was going to rain tomorrow. Therefore, it is going to rain tomorrow." On the face of it, this is as good as it gets, reasonable enough to accept, but, of course, fallible. But Groarke, in order to make it deductively valid, includes a hidden premise: "The weather network is never wrong." And he adds: "This is not, of course, a sound argument. The hidden premise is just silly." But why adding a *silly* premise? He speculates that the person who argues has a naïve confidence in the accuracy of the weather network's forecasts. Well, she might, but it is much more plausible to start with the everyday assumption that the arguer is using ordinary inductive type of reasoning. Groarke (2009: 98) considers this option in the form: "The weather network said it was going to rain tomorrow. Therefore, it is probably going to rain tomorrow." We are now supposed to add a hidden premise: "The weather network is *usually* accurate." And he thinks that a rational agent cannot believe in the first two premises without believing in the conclusion, so, given the premises, it *must* be the case that it will probably rain tomorrow. The main tenet of deductivism—that the truth of the conclusion of a good argument follows necessarily from the truth of the premises—is thus compatible with probabilistic reasoning.

I agree with Godden (2005: 173) that deductive standards preserve truth but not plausibility, probability, or likelihood. The lottery paradox is quite convincing: consider a fair 1000-ticket lottery that has exactly one winning ticket. For each *individual* ticket it is highly probable (99.9%) that it will not win, but we cannot *deduce* that it is highly probable that *no* ticket will win. A rational agent can believe in the whole lot of a thousand premises without believing in the deductively inferred conclusion.

Probability is a complicated issue, however, and a relation between deduction and induction is a huge issue (Jacquette (2009: 201, fn. 5) quotes a slogan attributed to Sellars: "An inference is either deductive or defective.") Still, I find it difficult to accept that the only good arguments are those for which *absolutely* no counterexample is to be found. Govier (1992: 403) offers a more plausible variety of *grounding* relations: premises ensure/entail/make it probable/support/give evidence … that the conclusion is true. Or, better still, in terms of counterexamples (Godden 2005: 171), accepting the premises of the argument, we should accept its conclusion if (i) the only counterexamples to be found are highly improbable; (ii) the only counterexamples to be found are less probable than the premises; (iii) no counterexample has been found yet (it has not been falsified); (iv) no counterexample is already to be found amongst our beliefs (coherence). In all of these cases it is logically possible for the conclusion to be false given the truth of the premises, but this alone does not automatically disqualify the inferential links in the arguments. Pluralism with relatively "high electoral threshold" so to speak (deduction, induction, perhaps conduction and analogical reasoning) seems to be the best option for the "following from" relation.

6.

Johnson and Blair (2002: 352) remarked that formal logic began with Frege as a revolution at the level of theory that later filtered down into logic textbooks. In informal logic developments at the theoretical level were largely motivated by the attempt to teach students how to assess arguments in use. We saw that deductivism offers an attractive toolkit. But there is another option. Suppose we take the bottom-up approach as our starting point for the general understanding of the "follows from" relation. One of the main logical skills (to be developed by "critical thinking courses") has always been the technique of counterexamples: the conclusion does not follow, it is *possible* to accept all of the premises but deny the conclusion. But one should consider *plausible* counterexamples only, not just any logical possibility. Weak deductivism already embraces arguments which are materially valid (it is *logically* possible for premises to be true and the conclusion false, but given the *meanings* of non-logical terms this is not possible). Why not continue in this spirit and impose further limitations on the range of possibilities to be considered?

Consider, as an example, some contemporary ecological hot issues in Slovenia. In a predominantly rural area with a high unemployment rate an international corporation proposed to build a car lacquering factory on mainly agricultural premises. Predictably a lively controversy ensued, the government and the defenders of the proposal argued in the following way:

> There is large unemployment and there are no other economic activities in this area, so we should not oppose the foreign corporation in their decision to build a car lacquering factory on these agricultural premises.

Is it possible for the premises to be true and the conclusion false? We are interested in serious, contextually relevant possibilities and the best way to focus on them would be to *extract* the "broad" logical form, something like:

> In the area A we need Y. Z is a source of Y. In the area A there are, currently, no other sources of Y. The benefits of Z outweigh the downsides. Therefore we should approve of Z.

We treat some of the *repeated* content expressions as variables and the rest of the argument as logical framework to be kept fixed when we engage in looking for potential counterexamples. Z (car lacquering factory) is really a source of Y (prosperity) in the area, but it is not the *only* possible source of prosperity and even if benefits outweigh the downsides it might still be sensible to deny the conclusion (just consider chemotherapy and cancer). Now consider a different argument based on the same pattern of reasoning. In a windy Karst area, rarely populated but otherwise a well-known bird resort, the government proposed to build wind farms. Again a lively controversy ensued:

> In the area A we need Y. Z is a source of Y. In the area A there are, currently, no other sources of Y. The benefits of Z outweigh the downsides. Therefore we should approve of Z.

The discussion was mostly about the *acceptability* of premises (opponents operate with a rather vague notion of downsides, including "degradation of the landscape" etc.) and it is again possible to accept the premises but deny the conclusion even if benefits outweigh the downsides. Here, it seems, given the "overall" damage done to the environment by other potential sources of electricity, this possibility is less relevant than in the first case.[2] Perhaps a purely deductive reconstruction is also possible—weak deductivism is an attractive option. One could add premises about the degradation of the landscape and the protection of birds on one side and new employments, less need for other, more problematic sources of energy on the other side and so on. But the list is not fixed, and it seems more plausible to incorporate the content of hidden premises as *guidelines* for potential counterexamples.

Aristotle already typically proves the invalidity of a given syllogistic mood by providing an argument displaying the given form but which is obviously *invalid* (with true premises and false conclusion). Cogency can be tested in the same way, by matching the structure of a given argument with that of an argument whose cogency is known or obvious. This tactic is called "refutation by logical analogy" and it is based on duplicating the core of an argument in another argument by varying certain inessential components (marked by variables) while preserving the essential ones. If the parallel argument is not cogent, the original argument is not cogent either. In classical logic the essential/inessential partition of vocabulary is given in advance, logical constants are essential, descriptive terms are variable. And, secondly, when inspecting the space of possibilities opened by the variable interpretations of non-logical constants, we have to consider *every* possibility. Not so when we search for counterexamples to cogency: a *limited* (relevant) set of interpretations has to be considered for 'A', 'Y' and 'Z'. According to Quine's formulation descriptive terms occur vacuously in logically val-

[2] So says the informal logician in the year of 2018. Interestingly enough, the car lacquering plant was actually built and windmills were not. As we all know, decisions are not always based on logic, even logic broadly understood.

id arguments and essentially in extra-logically valid arguments. But when considering *cogency* and testing for broad logical consequence in "natural" arguments some descriptive terms are contextually vacuous (replaced by variables *A*, *Y*, *Z* in our example) and others are fixed ('area', 'source', 'benefits,' …).

I think that this approach best captures the interplay of form and content, the mix of purely inferential and epistemological, so typical for "informal" evaluations. Adler rightly observes (1997: 335):

> The proper notion of structure or form is much broader than the notion of logical structure or form. Whenever we distinguish in an inference pattern between constant elements and variables, open to substitution, where the inference turns on the pattern of these constant elements, and not the substitutions for the variables, we are specifying a structure or form (Brandom 1988). Additionally, the pattern must yield a rich set of inferences. On this conception, criticizing some arguments for the falsity of a premise, when it expresses a rich, structural pattern, does constitute the finding of a defect in form.

Traditionally this broad notion of structure was associated with the shift from the *form* to the *matter*. Thus understood the form versus matter distinction relies crucially on a partition of the vocabulary: some of the terms of an argument are thought to pertain to its form, while others are thought to pertain to its matter. Logical constants remain fixed while substantial 'material' terms are replaced by schematic letters ("All A are B and all B are C, so all A are C") and the ruling out of true premises and a false conclusion is due to the meaning of *logical* terms. According to the material *consequence* the conclusion follows because of the meaning of non-logical terms. Bolzano speaks about the *deductive* consequence in the broad sense but I prefer to speak about the consequence in the *broad* sense (cf. Šuster 2012).

I think that the best contemporary development of this broader sense of form or broad consequence can be found in the work of David Hitchcock (2017). He first spoke about "enthymematic validity," then wrote about "non-logical consequence" and finally settled for "material validity" and "material consequence" in line with the established tradition. Material consequence is the relation that results when some but not all of the non-logical terms are treated as if they were logical. According to his definition (Hitchcock 2017: 124):

> A conclusion is a consequence of given premises if and only if the argument is an instance of an argument scheme, which may or may not be purely formal, that has no actual or counterfactual instances with true premises and an untrue conclusion, even though it has an instance with true premises and an instance with an untrue conclusion.

He later explains the inference-claim of an argument as the claim that it has a contentful covering generalization that is non-trivially true. A conclusion follows from stated premises in accordance with a counterfactual-supporting covering generalization of the argument's "associated conditional": the material conditional whose antecedent is the conjunction of the reasons and whose consequent is the claim. Freeman (2011: 176–179) nicely summarizes this approach in terms of a *recipe*. Consider:

Socrates is human. Therefore Socrates is mortal.

First identify the repeated content expressions in the argument and uniformly replace *repeated* content expressions with variables of the appropriate category (human, mortal):

x is human, therefore x is mortal.

The variable components are the ones such that "intracategorial" replacement of them results in an analogue which is a potential counterexample to the original argument. Now form the associated generalized conditional, the covering generalization (the conjunction of the premises of the argument as the antecedent and the conclusion as the consequent):

For every x: If x is human, then x is mortal.

To claim that the conclusion of an argument follows from the premises is, according to Hitchcock, to claim that the covering generalization is necessarily true for some sense of necessity.

The recipe might work for some simple arguments, but I think that ecological issues mentioned above already escape the purely "algorithmic" approach. Hitchcock rightly points out that in assessing whether any argument's conclusion follows from its premis(es), we regard certain components as fixed and others as variable. But in general we can only provide *guidelines* for determining which of the components are fixed and which are variable. Also, I can hardly agree with the *total* dismissal of deductivism: "The doctrine of implicit premises is largely a myth. Theorists of argumentation and practitioners of argument analysis and evaluation should abandon it" (Hitchcock 2002: 160). Some arguments should really be analysed as *enthymemes*, deductive patterns with missing premises. A principled division between *material* consequence and deductive consequence proper is still an open question (though Freeman 2011: 173–195, makes some interesting proposals).

In any case Hitchcock has developed a promising approach to understanding the "follows from" relation, and I cannot do justice to all of the details of his rich analysis. I think that *broad* logical consequence, based on the traditional idea of counterexamples and the interplay of form and content best captures the central idea of normative assessment in the area of everyday arguments, something like (Fisher 2012: 25): "Could the premises be true and the conclusion false judging by appropriate standards of evidence or appropriate standards of what is possible?"

7.

When explaining the "informal" terminology Blair (2015: 28) makes an interesting analogy:

> You need to be wary of the notion that in the term "informal logic," the word 'informal' means "informal" and the word 'logic' means "logic." It is like

> the use of the term 'football' north of Mexico. In the USA and in Canada, the games called "football" don't much call for the players to control a ball with their feet. Informal logicians use variables, and talk about argument schemes, which are quasi formal. So informal logic is not strictly-speaking informal. And if you understand by logic the study of axiomatized deductive systems, informal logic is not logic.

Let me further develop this analogy. According to *Wikipedia* "Football is a family of team sports that involve, to varying degrees, kicking a ball with the foot to score a goal. Unqualified, the word football is understood to refer to whichever form of football is the most popular in the regional context in which the word appears."[3] And even more formal *Encyclopœdia Britannica* characterizes football as "any of a number of related games, all of which are characterized by two persons or teams attempting to kick, carry, throw, or otherwise propel a ball toward an opponent's goal. In some of these games, only kicking is allowed; in others, kicking has become less important than other means of propulsion."[4] In the same spirit we could ask: are the boundaries of logic really determined by the rules of formalization, axiomatic systems and classical deduction? Theory of proofs, theory of models, recursive functions ..., belong to a certain "regional" variety of logic. But logic in a broad sense (patterns of reasoning which by a certain type of necessity preserve acceptability) can be played differently. True, the rules are not strict, but we play that game everywhere and every day.

References

Adler, J. E. 1997. "Fallacies Not Fallacious: Not!" *Philosophy and Rhetoric* 30: 333–350.

Adler, J. A. 2006. "Confidence in Argument." *Canadian Journal of Philosophy* 36: 225–258.

Baronett, S. 2015. *Logic (3rd).* Oxford: Oxford University Press.

Biro, J. and Siegel, H. 1992. "Normativity, Argumentation and an Epistemic Theory of Fallacies." In van Eemeren et al. (eds.). *Argumentation Illuminated*. Amsterdam: SicSat: 85–103.

Blair, J. A. 2012. *Groundwork in the Theory of Argumentation: Selected Papers of J. Anthony Blair*. Dordrecht: Springer.

Blair, J. A. 2014. "Informal Logic." In van Eemeren, F. H., et al. *Handbook of Argumentation Theory*. Dordrecht: Springer: 373–423.

Blair, J. A. 2015. "What is Informal Logic?" In van Eemeren, F. H in Garssen, B. (ed.). *Reflections on Theoretical Issues in Argumentation*. Dordrecht: Springer: 27–42.

Carroll, L. 1895. "What the Tortoise Said to Achilles." *Mind* IV ,14 (April 1895): 278–80.

Copi, I. M. and Cohen, C. 1990. *Introduction to Logic* (8th). NewYork: Macmillan.

[3] Cf. https://en.wikipedia.org/wiki/Football (accessed July 21th 2017).

[4] football. (2013). Encyclopædia Britannica. *Encyclopœdia Britannica Ultimate Reference Suite*. Chicago: Encyclopædia Britannica.

Copi, I. M., Cohen, C. and McMahon, K. 2014. *Introduction to Logic* (14th). Harlow: Pearson.

Cozzo, C. 2017. "Cogency and Context." *Topoi*. https://doi.org/10.1007/s11245-017-9462-z

Dummett, M. 1973. *Frege: Philosophy of Language*. Duckworth: London.

Eemeren, F. H. van. 2015. *Reasonableness and Effectiveness in Argumentative Discourse. Fifty Contributions to the Development of Pragma-Dialectics*. Dordrecht: Springer.

Eemeren, F. H. van, Grootendorst, R. 2004. *A Systematic Theory of Argumentation*. Cambridge: Cambridge University Press.

Eemeren, F. H. van. 2009. "The Study of Argumentation." In Lunsford, A. A., Wilson, K. H., Eberly, R.A. (eds.). *The SAGE Handbook of Rhetorical Studies*. SAGE Publications, Inc.: 109–24.

Feldman, R. 1994. "Good arguments." In Schmitt, F. (ed.). *Socializing Epistemology: The Social Dimensions of Knowledge*. Rowman & Littlefield: 159–188.

Feldman, R. 2014. *Reason and Argument* (2nd). Pearson Education Limited.

Fisher, A. 2012. "A Little Logic." In Ribeiro, H. J. (ed.). *Inside Arguments: Logic and the Study of Argumentation*. Cambridge: Cambridge Scholars Publishing: 21–36.

Freeman, J. B. 2011. *Argument Structure. Representation and Theory*. Springer.

Godden, D. M. 2005. "Deductivism as an Interpretive Strategy: A Reply to Groarke's Recent Defense of Reconstructive Deductivism." *Argumentation and Advocacy* 41 (3): 168–183.

Govier, T. 1992. "What is a Good Argument?" *Metaphilosophy* 23: 393–409.

Govier, T. 2010. *A Practical Study of Argument*. Belmont: Wadsworth.

Govier, T. 2018. *Problems in Argument Analysis and Evaluation*. Windsor Studies in Argumentation Volume 6. Windsor Ontario Canada.

Groarke, L. 1999. "Deductivism Within Pragma-Dialectics." *Argumentation* 13: 1–16.

Groarke, L. 2009. *An Aristotelian Account of Induction: Creating Something from Nothing*. Montreal and Kingston: McGill-Queen's University Press

Hanna, R. 2006. *Rationality and Logic*. Cambridge: The MIT Press.

Harman, G. 1986. *Change in View: Principles of Reasoning*. Cambridge: MIT Press.

Hintikka, J. 1999. *Inquiry as Inquiry: a Logic of Scientific Discovery*. Dordrecht: Springer.

Hitchcock, D. 2002. "A Note on Implicit Premises." *Informal Logic* 22: 159–160.

Hitchcock, D. 2017. *On Reasoning and Argument*. Dordrecht: Springer.

Hurley, P. J. 2015. *A Concise Introduction to Logic* (12th). Stanford: Cengage Learning.

Jackson, F. 1984. "*Petitio* and the Purpose of Arguing." *Pacific Philosophical Quarterly* 65: 26–36.

Jacquette, D. 2007. "On the Relation of Informal to Symbolic Logic." In Jacquette, D. (ed.) *Philosophy of Logic. Handbook of the Philosophy of Science*, Amsterdam: North Holland, Elsevier B.V.: 131–154.

Jacquette, D. 2009. “Deductivism in Formal and Informal Logic.” In Koszowy, M. (ed). *Informal Logic and Argumentation Theory*. Bialystok: University of Bialystok: 189–216.

Johnson, R. H. and Blair, J. A. 1977. *Logical Self-Defense*. Toronto: McGraw Hill-Ryerson.

Johnson, R. H. and Blair, J. A. 2002. “Informal Logic and the Reconfiguration of Logic.” In Gabbay, D. M., Johnson, R. H., Ohlbach, H. J. and Woods, J. (eds.). *Handbook of the logic of argument and inference*. Amsterdam: North Holland, Elsevier B.V.: 339–396.

Johnson, R. H. 2000. *Manifest Rationality*. London: Lawrence Erlbaum Associates.

Johnson, G. 2016. *Argument and Inference*. Cambridge: The MIT Press.

Mendelsohn, R. L. 1996. “Diary: Written by professor Dr Gottlob Frege in the time from 10 March to 9 April 1924.” *Inquiry: An Interdisciplinary Journal of Philosophy* 39 (3–4): 303–342.

Monk, R. 2017. “Gottlob Frege: The machine in the ghost.” *Prospect Magazine*, October 2017. https://www.prospectmagazine.co.uk/philosophy/the-machine-in-the-ghost (accessed February 23th, 2018).

Perelman, C. and Olbrechts-Tyteca, L. 1971. *The New Rhetoric*. Notre Dame: University of Notre Dame Press.

Pinto, R. C. 2001. *Argument, Inference and Dialectic*. New York: Springer.

Plumer, G. 2016. “Can Cogency Vanish?” *Cogency: Journal of Reasoning and Argumentation* 8 (1): 89–109.

Quine, W. V. O. 1950. *Methods of Logic*. Cambridge: Harvard University Press.

Quine, W. V. O. 1986. *Philosophy of Logic* 2nd. Cambridge: Harvard University Press.

Sosa, E. 2004. “Relevant Alternatives, Contextualism Included.” *Philosophical Studies* 119: 35–65.

Šuster, D. 2012. “Informal logic and informal consequence.” In Trobok, M., Miščević, N. and Žarnić, B. (eds.). *Between logic and reality: modeling inference, action and understanding*. Dordrecht: Springer: 101–120.

Vorobej, M. 2006. *A Theory of Argument*. Cambridge: Cambridge University Press.

Walton, D. N. 1991. *Begging the Question: Circular Reasoning as a Tactic of Argumentation*. New York: Greenwood Press.

Walton, D. N. 1996. *Argumentation Schemes for Presumptive Reasoning*. New York and London: Routledge.

Walton, D. N. 2006. “Epistemic and Dialectical Models of Begging the Question.” *Synthese* 152: 237–284.

Walton, D., Reed, C. and Macagno, M. 2008. *Argumentation Schemes*. Cambridge: Cambridge University Press.

Woods, J. 2004. *The Death of Argument. Fallacies in Agent Based Reasoning*. New York: Springer.

Wright, C. 2002. “(Anti-)Sceptics Simple and Subtle: G. E. Moore and John McDowell.” *Philosophy and Phenomenological Research* 65 (2): 330–348.

Croatian Journal of Philosophy
Vol. XVIII, No. 54, 2018

Informal Reasoning and Formal Logic: Normativity of Natural Language Reasoning[1]

NENAD SMOKROVIĆ
University of Rijeka, Rijeka, Croatia

Dealing with deductive reasoning, performed by 'real-life' reasoners and expressed in natural language, the paper confronts Harman's denying of normative relevance of logic to reasoning with a logicist thesis, a principle that is supposed to contribute for solving the problem of incongruence between descriptive nature of logic and normativity of reasoning. The paper discusses in detail John MacFarlane's (2004) and Hartry Field's (2009) variants of "bridge principle". Taking both variants of bridge principles as its starting point, the paper proceeds arguing that there is more than one logical formalism that can be normatively suitable for deductive reasoning, due to the fact that reasoning can assume different forms that are guided by different goals. A particular reasoning processing can be modelled by specific formalism that can be shown to be actually used by a real human agent in a real reasoning context.

Keywords: Logic, real-life reasoning, normativity, deductive reasoning.

1. *Introduction*

The paper deals with the normativity of reasoning, specifically with *deductive reasoning*, performed by 'real-life' reasoners and expressed in natural language. Deductive reasoning in a 'real-life' situation might seem as a kind of oxymoron. If reasoning is deductive it seems that it should be in accord with the rules of deductive logic. As it is well empirically documented, everyday reasoning can hardly satisfy deductive logic's standards. The question of normativity I am interested in is whether formal logic, or at least a kind of formal logic, can still have a decisive

[1] The writing of this paper was supported by the Croatian Science Foundation under the grant number IP-06-2016-2408.

normative implication for reasoning. Why is this question important? It is important due to the fact that a drastic denial of normative impact of logic to reasoning leaves us without the safe criteria of normativity. We are in this case left only with the appeal to intuitions that are supposed to be the arbiter of correctness of reasoning. On the other hand, if there is a plausible theoretical connection between logic and reasoning, whereas logic can also be of non-classical, even non-monotonic kind, our understanding of reasoning will be on a much firmer ground.

The problems for applying normativity of logic to reasoning, in the formal setting, are in their sharpest form stated by Harman (1986). He famously proclaimed the independence of logic to reasoning arguing that there is a huge gap between logic *that describes the relation of implication* and the normativity of reasoning that has to do *with what we should believe*. However, this paper is defending the *logicist thesis*. The logicist thesis is the claim that there is, to use MacFarlane's formulation: "some connection between logical validity and the evaluation and criticism of reasoning" (MacFarlane 2009: 2). In other words, general logicist thesis proclaims that logic (it needn't be classical logic, even not one of the necessarily truth preserving kind of logic) has a decisive normative role for reasoning.

In §2 the paper starts with some remarks on reasoning, particularly concerning the difference between deductive reasoning and deductive logic. This difference certainly justifies Harman's denying the normative role of logic for reasoning. Nevertheless, a number of philosophers have recently put forward their versions of normativity of logic opposing Harman's view. Let me mention some of them: J. MacFarlane (2004), Hartry Field (2009), Peter Milne (2009), Caterina Dutilh Naves (2013, 2015). They want to answer Harman's challenges articulating what I call *logicist thesis* in the form of different versions of *bridge principle*, a principle that is supposed to contribute to solving the problem of incongruence between descriptive nature of logic and normativity of reasoning. The paper discuses in some detail John MacFarlane's (2004) and Hartry Field's (2009) variants of "bridge principle". They both take for granted that on the one side of the bridge there is a particular valid logical form (MacFarlane takes it to be classical logical validity while Field allowed different kinds of logical validity) and on the other one, more or less uniform, deductive behaviour that is to be normatively captured by proposed formalism. However, contrary to them, I'm proposing the picture of deductive reasoning that manifests itself in different forms, each of which can be modelled by a different logic.

In §3 the concept of normativity will be considered. I will tackle the general question of the role of normativity in researching reasoning and in more details the issues of the scope of applicability of normative rules and of the ways in which the normative impact of logical rules on reasoning can be understood. Concerning the first issue, I'm embracing the view that norms can be applicable to those who apprehend them, while regarding the second issue I advocate the view that logic can be normative in a stronger sense, as a *guidance for reasoning*. It contests the thesis promoted by Ferrari and Moruzzi (2017) that only the weaker sense of normativity can have the normative role in logic, claiming that normative rules are mere *criteria of correctness*.

My proposal, in §4, concerning MacFarlane's and Field's Bridge principles is that more than one logical formalism can be normatively suitable for deductive reasoning due to the fact that reasoning can assume different forms that are guided by different goals. Namely, reasoning shows up in a variety of forms. It arises in everyday argumentations and debates aiming at a kind of shared agreement, but also appears in other contexts such as juridical debates or in scientific, philosophical, even mathematical dialogues. In each of these contexts reasoning might have different goals. The goal of proving the theorem is different from the goal to show that an accused is guilty beyond any reasonable doubt, which is, again, different from the goal to make understand one that the bus will start from platform 1 when it is so stated in timetable and no other information is available. Each of these reasoning forms can be captured by suitable logics.

Let me now indicate what I mean by a *form* of reasoning. It is an inference form that is relevant for a particular real-life situation in the sense that this form is **just** sufficient for achieving a particular goal. As Varga, Stenning, Martignon, (2015: 1) put it, "computational efficiency is an opportunity cost of expressive power". This form of reasoning is normatively justified if this form can be connected with a kind of validity that the thinker can apprehend or recognize as valid.

2. *Remarks on reasoning and Harman's objections to normative role of logic*

a) *Deductive reasoning and deductive logic*

By deductive reasoning I mean a process of reasoning that guarantees a transition of the truth from a set of propositions, believed or known by the agent, to the conclusion. Let me illustrate the process by an example of a reasoner who, from the beliefs:

> *The 8 am bus from Rijeka to Zagreb starts either from platform 1 or from platform 2,*
> *The bus does not start from the platform 2,*

infers to the conclusion

> *Therefore, it starts from platform 1.*

This piece of reasoning is a subject to assessment. It is a correct reasoning. Talking about correctness or goodness of the episode of reasoning we inevitably invoke the normativity dimension (consider either as the first or as the third person perspective) of reasoning. However, as it is well known, normativity, particularly normativity of deductive reasoning, is a highly contentious topic. We will briefly tackle some of the issues. The first one is the relation between deductive reasoning and deductive logic. To say that deductive reasoning preserves the truth in inferring from the premises to the conclusion is to indicate, in one way or the other, that the normative standard for deductive reasoning is *deductive logic*. By deductive logic many logicians and psychologists mean a logical calculus that *necessarily preserves the truth,* notably, classical predicate logic[2] (CPL, henceforth). CPL by definition is extensional and truth-functional. Valid reasoning in this sense is represented by the argument in which conclusion is a logical consequence of its premises expressed as: *if all its premises are true, conclusion can't be untrue.*[3] Deductive reasoning that as valid is determined by basic properties of classical logic: monotonicity and necessarily truth preservation.

In accord with this line of thinking, deductive reasoning in natural language is deductively valid if it can be correctly translated into an argument that is semantically valid in a formal system (notably, CPL). We can consider this formulation as a standard view of deductive reasoning. This view presupposes two things: it equated the deductive reasoning with deductive logic, and further, it equated logical validity with the necessarily truth preservation.

The problem with the view that deductive reasoning is equal to deductive logic (that implies that the notion of logical consequence is CPL notion) is that reasoning performed in natural language is not syntactically or extensionally valid but at best intentional (semantically valid). Reasoning in natural language, in contrast to an argument form expressed in formal language, is sensitive to propositional content that should be interpreted in connection to the real world. In this interpretation people's knowledge of the world and evidence they have play an important role in their reasoning (what is irrelevant in the formalized classically valid argument). The real-life reasoning in natural language, therefore, hardly satisfies properties of deductive logic. The inferences performed in this domain are hardly *necessarily* truth preserving. Even more, they are often non-monotonic.

Having formulated the difference between reasoning and logic, the crucial issue of the paper becomes visible, namely, can we, in spite of the described characteristics, consider everyday reasoning as deductive? Many would say that, in so far, if it is not classically logically valid it is not deductive either. We are here faced with the dilemma: either real-life reasoning is not deductive, or deductive reasoning is to be weakened and broadened in a sense.

[2] Due to the limitation of the paper I'm neglecting the view held by in no way marginal number of logicians that see intuitionistic logic in the position of *the* logic that necessarily preserves truth.

[3] According to Tarski, logical consequence should be understood in terms of necessarily truth-preservation (Tarski 1956: 411), which, in turn, can be sharpened model-theoretically as follows: a sentence p follows logically from a set of sentences S just in case every model of S is a model of p (Tarski 1965: 417).

Deductive reasoning can be weakened so that it can be modelled by formal systems, other than CPL that possibly suits better real-life reasoning's salient characteristics. Here is one of those characteristics: it is often the case that real-life inferences are not classically valid, in the sense that *if all premises are certain, so is the conclusion*, but instead, (at least) some premises are *probable in a various degree*. Suppose, as Hayek (Hayek 2001) invites us to suppose, "that we want the probability of a conclusion of a given valid argument to be above a particular threshold". The answer to this question can be given through *probability logic* that is "the study of the transmission (or lack thereof) of probability through inferences" (Hayek 2001).[4] In this logic the traditional concern with the truth of premises is replaced with the concern about their probabilities. Such logic is certainly deductive, although non-monotonic (initially assigned degree of probability to the conclusion may later be retracted in the face of a new evidence) and not strictly truth-functional.

The other salient characteristic is that in everyday situations a conclusion from a given set of premises is often reached *defeasibly*. It means that the real-life reasoner reserves the possibility to *retract* from the originally reached conclusion in the light of new information or adding a new proposition to the original set of premises. This characteristic can be modelled by different variants of *default logic*.

Coming back to our dilemma, Gilbert Harman, supposing that deductive logic is equal to classical logic, is the leading authority of the view that reasoning does not correspond to deductive logic. In so far they are distinct. Logicists hold **a** different stance.

b) *Harman's objections to normative role of logic*

Let me outline the alleged difference between the descriptivism of logic on one side and the normativity of reasoning, on the other, posed by G. Harman (1986). According to this, logic merely *describes* logical relations; it does not *prescribe* what we should believe. For example, logicians *describe* an argument as a valid saying that it is impossible for the premises to be true without the conclusion to be true. Their main interest is in the relation between propositions and in what follows from what. There is nothing normative in this claim.

In a nutshell, Harman's reasons for divorcing logic from reasoning are:

> *Objection from belief revision*: claims about logical validity are not explicitly normative in their content. They do not tell us what we should believe. If, for example, one believes *p* and believes *p implies q* and recognizes them to jointly entail *q*, one is not under any particular normative obligation to believe *q* (for instance, if q is at odds with one's other beliefs, it would be unreasonable to accept *q*).

[4] In his article Hayek (2001) presents the general features of Adams' probability logic (1998), although other systems of probability logic are at stake.

> *Clutter avoidance*: There is a worry about "clutter avoidance" (is one really obliged to believe all of the infinitely many trivial logical consequences of one's beliefs?).
> *Excessive demands*: Norms of logic (might be) so demanding that no human being could possibly satisfy them. Namely, due to the limitations of cognitive resources and computational powers, no one can believe all consequences of his beliefs.

The *normative claim*, in contrast to merely logically descriptive one, has to do with thinkers (actual or potential) that perform the inferences, with the goals their reasoning process aimed to and with the doxastic states engaged in the process. The normative counterpart of the above descriptive example might be expressed in this way: in order to be *rational*, a *reasoner* (actual or potential) *should*, if she *believes* (or accepts) a set of propositions and *believes* that the conclusion follows from the premises, *believe* or *accept* the conclusion. At any rate, there is a significant difference between the rules of logic (or logics) and its normative counterpart. The normative claim, in contrast to descriptive ones, has to mention a reasoner's goal, her doxastic state and the particular deontic operator.

c) *Logicists' answers*
MacFarlane's bridge principle

In spite of the mentioned difference between logic and reasoning, in everyday reasoning processes reasoners tend to preserve the truth of the premises in the conclusion (although the truth preservation need not be necessary, as we will see in Field's formulation), and hence to obtain the deductive character of the informal reasoning. The logicists, in order to meet Harman's challenges, aimed to connect the formal logical consequence (or validity or implication relation) with the informal understanding of consequence in the way that formal consequence can be normative for informal reasoning.

The logicists hold that logical validity on one side and how we ought to think, on the other, should and can be connected. In this sense MacFarlane says:

> Why do we bother studying this notion (validity) at all? Surely it is because we think there is some connection between logical validity and the evaluation and criticism of reasoning. If we could get clearer about this connection, we could transpose questions about logical validity into questions about how we ought to think. (MacFarlane 2004: 2)

To meet Harman's challenge John MacFarlane has meticulously proposed the way to establish a connection articulated as a *bridge principle* able to override the gap between logical system that is descriptive and the normativity of reasoning (2004). Its goal is to "transpose questions about logical validity into questions about how we ought to think" (MacFarlane 2004: 2). Just to indicate the idea, the bridge principle (BP) is a *material conditional* that connects a valid logical form, say: A, B (as antecedent) and a normative claim that is compound of thinker's doxastic states (S believe, S knows) and deontic operators as *should*, is *permit to* or *has a reason* (as consequent). The conditional asserts that, *if* there is a *valid* logical form *then* comes the normative claim: (for instance) if S *believe* that P_1, ..., P_n together imply Q*, then S ought* to *believe* that Q. Formally: *'If P_1, ..., $P_n \vDash Q$, then* Φ' where Φ is a normative principle. MacFarlane's general strategy is to hold fixed *classical logical formulation of validity* as antecedent and combine elements in normative claim (consequent) in order to get the most "natural" and "realistic" combination of the mentioned parameters, for which the classical validity can play a normative role, avoiding in this way Harman's objections.

To obtain this, MacFarlane combines various types of *deontic operators* (strict obligation, permission or defeasible reasons for belief), the scope of the deontic operator in the conditional (narrow or wide: does deontic operator govern the consequent of the conditional, both the antecedent and the consequent or the whole conditional?) and doxastic states (believing and knowing).

Let me present in somewhat systematic way the parameters used in determining the normativity claim of the bridge principle:

(1) *Deontic operators*
(o) 'Ought'/obligation
(m) 'May'/permission
(dr) Defeasible (pro tanto) reason.

(2) *Polarity*
(+) Positive (o), (m) or (dr)
(–) Negative (o), (m) or (dr)

(3) *Scope of the deontic operator—e.g. 'o' denoting 'ought'*

Narrow scope: (n) (if P, then o (Q))
Wide scope: (w) o (if P, then (Q))

Governing both the antecedent and the consequent of the conditional: (b) o (if P, then o (Q)).

To complete the parameters the doxastic states of the subject are to be added. Namely, doxastic *restrictions* can be imposed on the antecedent part of the principle, in the sense that subject knows, apprehends or recognizes that particular form is logically valid. If such restriction *is* imposed the principle takes subjective or internal reading, contrary to objective reading when such restriction is not imposed. Although the combination of all parameter settings gives 36 bridge principles in total, I will illustrate MacFarlane's idea with four examples using only 'ought' operator, positive polarity, narrow and wide scope and doxastic state "know".

> (Narrow scope): If A, B ⊨ C, then if you believe A and you believe B, you ought to believe C.

> (Narrow scope + ‘know’): If you know that A, B ⊨ C, then if you believe A and you believe B, you ought do believe C.
>
> (Wide scope): If A, B ⊨ C, then you ought, if you believe A and you believe B, you believe C.
>
> (Wide scope + ‘know’): If you know that A, B ⊨ C, then you ought, if you believe A and you believe B, you believe C.

MacFarlane chose wide scope + ‘know’ formulation as the appropriate form for the normative claim. Nevertheless, his final decision for BP slightly changes the above formulations giving to BP an even stronger subjective reading. He takes the logically valid *schema* instead of classical logical consequence to figure in the position of the antecedent. The stronger subjective or internal note is given in the formulation that the subject knows that schema S is valid and, furthermore, the subject apprehends the given inference as an instance of S. The formulation is:

> If [you know that] the schema S is formally valid and you apprehend the inference A, B / C as an instance of S, then (normative claim about believing A, B, and C). (MacFarlane 2004: 22)

Eventually, MacFarlane gives the final form to BP, which I will take as his definite stance:

> If schema S is formally valid and you apprehend the inference A, B / C as an instance of S, then you ought to see to it that if you believe A and you believe B, you believe C. (MacFarlane 2004: 24)

Hartry Field’s variant of BP

In (Field 2009) Hartry Field developed his view on normativity of logic and offered his variant of BP. Unlike MacFarlane, he introduces *degrees of beliefs* expressed in the notion of probabilities, as doxastic units, instead of *full beliefs*. On the other side, similar to MacFarlane, he gives a subjective reading to the formulation of BP. Let’s take a closer look at his variant of BP:

> If it’s obvious that A_1, ..., A_n together entail B, then one ought to impose the constraint that P(B) is to be at least P(A_1)+ ... +P(A_n)−(n−1), in any circumstance where A_1, ..., A_n and B are in question. (Field 2009: 259)

Subjective reading is evident in the formulation “if it is obvious”, where *obvious* is to be understood as agent-relative. Obviousness as a doxastic restriction on the implication relation “A_1, ..., A_n together entail *B”* is equivalent to MacFarlane’s use of the notion of subject’s “apprehension” that the inference is an instance of the schema. Again, in contrast to MacFarlane, the relation “A_1, ..., A_n obviously together entail *B”* is not understood exclusively as a classical logical relation (material implication).[5] Field allows here the pluralistic reading. He says:

[5] It is not clear whether MacFarlane himself persists on material implication in his final formulation. It is the fact that at the certain point in his paper he changes the notation and A, B ⊨ C replaces with A, B / C that indicate that the relation is weaker than the material conditional.

"Whatever logic is assumed correct, it seems to me that

> if B is obviously entailed by A in that logic, a proponent of that logic should believe B to at least as high degree as A"

Let me stay a bit longer at Field's understanding of the relation "A_1, ..., A_n together entail B". We already see that the entailment relation or, if you prefer, the consequence relation, need not be classical since the plurality of logic is allowed. As the necessary truth preservation (NTP) is a substantial feature of the classical logical consequence, does Field allow a consequence relation that is not NTP? As a matter of fact, one of Field's important claims is that the relation of logical consequence *is not* the relation of the *necessary* truth preservation. In his own wording:

> I'm inclined to state my conclusion by saying that the validity of a rule does not require that it generally preserve truth. However, some may think that this simply violates the meaning of the term 'valid': 'valid', they may say, simply means 'necessarily preserves truth', or 'necessarily preserves truth in virtue of logical form'..." (Field 2009: 266)

Instead of defining validity in terms of NTP, he proposes:

> Perhaps we should redefine validity, not as (necessarily) preserving truth in general but as (necessarily) doing so 'when it matters'? (Field 2009: 266).

And finally:

> I basically said that a rule 'preserves truth when it matters' if it preserves truth *when applied to premises that can be established or are rationally believable.* (Field 2009: 266)

Comparing two variants of BP, MacFarlane's and Field's, I would say that both of them successfully connect the formal logical consequence with its informal understanding in reasoning, providing in this way the normative standard for reasoning. Still I take that Field's variant suits my purposes better. MacFarlane's formulation of normative rules requires from the agent to make only those inferences that he apprehends as instances of the valid schema. Still, determining a schema as classically valid, MacFarlane requires that it necessarily preserve the truth. In so far, normatively correct inferences are those that are necessarily truth preserving. Field's variant of BP is more liberal, allowing the implication (consequence) relation that is *not necessarily* truth preserving, which is much closer to the real-world reasoning that tends to preserve the truth but usually only "when it matters".

This line of thinking fits well with my proposal claiming that more than one logical formalism can be normatively suitable for deductive reasoning. The idea is that people in real-life situations perform different forms of reasoning, each form guided by a different goal. Being engaged in various forms, accomplishing different goals, they can be normatively warranted from the viewpoint of different logics.

3. *Normativity of reasoning*

In so far we have been discussing the normative role of logic in reasoning. In this relatively short part of the paper we will first make a brief remark on the general question of the role of normativity in reasoning (whether expressed in logical rules or somehow differently). After that, we are going to discuss two important issues concerning normativity. The first one regards the scope or domain to which normative rules can be applicable, while the second one considers different ways in which rules of logic can be normative.

Concerning the first question of the place and role of normativity in examining reasoning, there is a tendency in recent writings to thoroughly eliminate the role of normativity in investigating reasoning. It is in this vein that Elquayam and Evans (2011) advocate the idea of a complete abandoning of normativity in the psychological scientific practice. As it is hard to see good reasons for such a claim[6], I am starting from the opposite view. Concerning the role of normativity in reasoning, the claim is that the very concepts of reasoning, argument and argumentation are entirely normative. This is obvious in the scientific field as well as in the everyday social intercourse. In all kinds of discourse people are prone to recognizing a chain of reasoning as a 'good' one and an argument as a 'correct' one. They do this from the first-person perspective, associating a *degree of confidence* to the correctness of their judgments and other outcomes of reasoning processes. People continually do this also from the third person perspective, assessing reasoning of others as correct or incorrect. In empirical investigations of reasoning, "without norms of some kind, we cannot interpret the data participants produce" (Achourioti, Fugard, Stenning 2014). Therefore, I take for granted the inevitability of normativity in reasoning.

Although the host of issues and open questions concerns the area of normativity, we will tackle two of them, namely, what can be the domain of application and how to understand that rules are applicable to subjects.

The question of *domain or scope of applicability* can be formulated in the following way:

Are normative rules of reasoning applicable universally to the wide domain of all rational beings, or are normative rules specific, having a domain of application only to those who apprehend or understand applied normative rules? Relative to the latter disjunct, normative rules have a restricted application relative to the subject's apprehension of the rule.

Concerning the former understanding of normativity, this approach has often been put forward in the traditional but also in the recent literature. The problem with this approach is, I hold, in the aprioristic

[6] Due to the tolerable length of the paper, I am not able to support my judgment with the extended argumentation as that would deserve a separate paper.

determination of normative rules and in the generality its nature is determined. A typical example of such a consideration is Frege's claim: "Logic prescribes universally how one ought to think if one is to think at all" (Frege 1893). In this way, reasoning rules proclaimed as normative are quite general logical principles that are understood as belonging to CPL (for example: reasoning ought to be consistent). At the same time, it is this understanding of normativity that underlines Harman's objections regarding the connection between logic and reasoning. On the other side, normativity as restricted to the subject's apprehension seems to be at the basis of MacFarlane's and Field's approaches. I'm siding with this latter, narrow or restricted, view.

The second question concerns the possible ways in which logic can be normative for reasoning. Recently, Florian Steinberger (Forthcoming) distinguished three ways in which this question can be understood.

According to the *first one*, normative rules are supposed to prescribe *directives* for reasoning in the sense that they have a guiding role (from the subject's, first person, perspective) in deciding what to believe.

According to the second one, they are supposed to give the criteria or standards for the *evaluation* of the *good* reasoning (from the third person perspective).

Finally, normative rules might play a role of the third personal *appraisals* by which one can blame or praise an agent for her inferential conduct.

For the purposes of this paper it is sufficient to consider only the first two roles the normative rules can play; let me call them *directive* and *evaluating* roles. Assigning the *directive* role to normative rules for reasoning one understands normativity in a stronger sense than taking it to have only evaluative role. If normativity is directive it is in principle also evaluative, while the evaluative role does not imply the directive one. It seems that Harman had in mind the directive role of logic for reasoning when he denied its normative influence. Accordingly, in order to defend the *normativity thesis* against Harman's objections, the strong, directive meaning of normativity has to be embraced.

Summing up the discussion in this chapter and putting together the questions of scope and of ways of understanding normativity, among the possible answers to these questions I'm picking up the *restricted*, apprehensive scope of rules' application and the *directive,* guiding role of normative rules. They together determine the desiderata, for, I hope, a promising way to uphold my view of normativity that is going to be exposed in §4.

4. *Forms and norms of reasoning*

The goal of the *normativity thesis* I'm supporting is to uphold a tighter connection between the normativity expressed in logical formal rules and the pre-theoretic comprehension of logical principles used in actual reasoning. MacFarlane and Field have been formulating variants of bridge principle that have in common the subjective, restricted understanding of rules applicability. I take it to be the significant desideratum of normativity to which I'm adding the directive role of normativity. These two desiderata of normativity, to which I will refer thereof as to *restrictiveness*, and to *directiveness,* can make possible a promising step forward in this direction. *Restrictiveness* and *directiveness* are obviously connected. Logical rules can have a guiding role for those who are able to apprehend them in a certain sense. Expressing the same thing in different way, we can say that only those who have the rule represented in explicit or implicit way can *follow* a rule. Otherwise, the agent's inferential behaviour can be only evaluated from the third person perspective.

Starting from *restrictiveness*, it is an open question in what sense the apprehension of logical rules is to be understood. The view that apprehension should be understood as an explicit mastering of the rule is clearly over-demanding and should be, therefore, ruled out as a candidate. As a promising approach to the answer I take MacFarlane's stance that to apprehend an inference is to see it as having a certain logical structure. But he claims more than that. He claims that:

> On this view, all logical norms have their source in the thinker's "apprehension" of inferences as having a certain formal structure. (MacFarlane 2004: 22)

And in clarifying in what sense apprehension is to be understood, he says:

> My own view is that apprehension should not be intellectualized to the extent that it requires a completely explicit understanding of what an inference schema is, the kind one would get from an encyclopedia article on the subject. It is something more basic than that. But it is important that apprehension be something for which one can take responsibility and give or receive criticism. (MacFarlane 2004: 22).

I'm in accordance with this view on apprehension. Still, it is noteworthy to make some caveats regarding this formulation. Let me start with *taking responsibility and giving or receiving criticism*. This formulation seems to mark what it means that apprehension is *more basic* than explicit understanding. According to this, one apprehends an inference as an instance of inference schema (IS) if one is responsible in the sense that one intends to infer according to IS. One is responsible in this sense for all and only episodes of reasoning that she apprehends as belonging to IS. It goes without ado that IS itself should be valid. But let's note that agent's apprehension has no role in recognizing an IS as valid. Although it is not quite clear whether MacFarlane considers the validity of IS independent of agent's apprehension, it seems as he holds that IS's validity is fixed as necessarily truth preserving (NTP). Accordingly, an agent is normatively responsible for an instance of inference if she apprehends that it belongs to IS, but the kind of required validity for IS is fixed as NTP. Let me call this approach *apprehension plus fixed IS.*

Although it is supposed that this approach can challenge Harman's objections, it seems that it can't meet all of them. It is particularly vulnerable to *objection to belief revision*. Let me use for illustration the example from §2. Here we had a reasoner who, looking at the timetable, comes to know that:

1. The 8 am bus from Rijeka to Zagreb starts either from platform 1 or from platform 2,
2. *The bus does not start from the platform 2,*

and from that she infers to the conclusion

3. Therefore, it starts from the platform 1.

Let's remind that this is an instance of the real-life reasoning where premises sometimes can't be taken with absolute certainty (*reasoning in uncertainty*) or a new evidence can produce the contradiction, for example, the added information that platform 1 is at the moment unavailable. In this case, the agent is faced with contradictory beliefs. If we are trying to model her reasoning in the frame of classical logic, the reasoning is valid even when the reasoner makes whatever conclusion (in accord with the principle *ex falso quodlibet*). Harman takes such a situation as an evidence for separating logic from reasoning (Harman 1986). When new information is added, our agent is forced to abandon her premise 1, but in this case logic does not guide or even recommend any action.

Coming back to MacFarlane BP, when validity of IS is equated with the necessary truth preservation, the *apprehension plus fixed IS* can't solve the problem. But, if we consider other kinds of validity grounded in different logics, including non-monotonic ones (notably probabilistic and default logics) that better suit the real-life reasoning, the solution seems to be more probable. The employment of a particular kind of default logic could be especially suitable in our example. Varga, Stenning, and Martignon (2015) have proposed *closed world semantics,* which is a variant of default logic. Closed world assumption provides a valid, truth-preserving inference that is represented with this conditional (Varga, Stenning, and Martignon 2015: 3):

$$p \,\&\, {\sim}ab \rightarrow q$$

meaning: *If p and nothing abnormal is the case, then q.*

In the situation as the above mentioned, an agent can apprehend: p $\&\ {\sim}ab \rightarrow q$ as a valid inference schema and in addition apprehend the episode of reasoning:

> *The 8 am bus from Rijeka to Zagreb starts either from platform 1 or from platform 2,*
> *The bus does not start from the platform 2,*
> *Therefore, it starts from the platform 1,*

as an instance of this schema.

This consideration nicely fits Field's proposal of the BP as closer to the solution we are looking for. This approach is more liberal than MacFarlane's in regard to the possible kinds of validity of inference. Allowing the plurality of logics,[7] Field's proposal makes it possible to model also those forms of real-life reasoning that fall outside the scope of necessarily truth preserving arguments. Particularly it is possible with reasoning in uncertainty (can be modeled by probabilistic logic) and with defeasible reasoning (can be modeled by default logic).

The proposal of apprehension of both the validity of IS and the validity of instance of reasoning as belonging to apprehended kind of validity corresponds to the view of reasoning appearing in different forms of reasoning. Reasoning as a cognitive activity is not a uniform endeavor and it can't be idealized as having a closed list of characteristics and normative constraints. On the contrary, as it is indicated above, people in real-life situations perform different forms of reasoning, each form guided by a different goal. Being engaged in various forms, accomplishing different goals, they can be normatively warranted from the viewpoint of different logics.

The relevance principle tells us that people economize with their cognitive resources. As Varga, Stenning and Martignon put it "computational efficiency is an opportunity cost of expressive power" (2015: 1). There are some goals a thinker can obtain mobilizing mostly his implicit deductive inferential performance, while for other goals the explicit, reflective thinking will be necessary. The goal of proving the theorem is different from and requires different cognitive effort than the goal to show that an accused is guilty beyond any reasonable doubt, which is, again, different from the goal to make understand one that the bus will start from platform 1 when it is so stated in timetable and no other information is available. Each of these reasoning forms can be captured by suitable logics.

However, for any form and goal of deductive reasoning there is an adequate normative system that can direct this reasoning toward the "rational" achievement of the goal. Which kind of logic is to be employed as normatively relevant for a particular form of reasoning is partly an empirical question. I am proposing the approach to the nor-

[7] The idea of the plurality of logic I have in mind is quite close to Beall and Restall's theory (2006). They consider any logic whose notion of validity satisfies what they call *Generalized Tarski thesis*, GTT.

GTT: "An argument is valid$_X$ in every case$_X$ in which the premises are true, so is the conclusion."

Variable x ranges over types of cases. Shapiro (2014) clarify the relation of logics, validity and cases as follows:

> "Classical logic results from GTT if 'cases' are Taskian models; intuitionistic logic results if 'cases' are constructions or stages in constructions (i.e., nodes in Kripke structures); and various relevant and paraconsistent logics results if 'cases' are situations (of a particular sort). In present terms, then, Beall and Restall take logical consequence to be folk-relative to kinds of cases. In their view, for example, the low of excluded middle is valid relative to Taskian models, invalid relative to construction stages (Kripke models); and the argument of *ex falso quodlibet* is valid relative to Tarskian models (and possible worlds), invalid relative to situations." (Shapiro 2014: 33).

mativity looking at different logical formalisms on the one side and on the other the actual human reasoning behaviour, adjusting one to the other through a kind of reflexive equilibrium.

References

Achourioti, T., Fugard, A., Stenning, F. 2014. "The empirical study of reasoning is just what we are missing." *Frontiers in Cognitive Science* 5: 1159, doi: 10.3389/fpsyg.2014.01

Adams, E. W. 1998. *A Primer of Probability Logic*. Stanford.

Beall, J. C. and Restall, G. 2006. *Logical Pluralism*. Oxford: Oxford University Press.

Dutilh Novaes, C. 2013. "A Dialogical Account of Deductive Reasoning as a case Study for how Culture Shape Cognition." *Dialectica* 96 (4): 587–609.

Elquayam, S. and Evans, J. 2011. "Subtracting 'ought' from 'is': *Descriptivism versus normativism in the study of human thinking*." *Behavioral and Brain Sciences* 34: 233–290.

Ferrari, F. and Moruzzi, S. 2017. "Logical Pluralism, indeterminacy and the normativity of logic." *Inquiry*, doi: 10.1080/0020174X.2017.1393198

Field, H. 2009. "What is the Normative Role of Logic." *Proceedings of the Aristotelian Society Supplementary Volume* lxxxiii.

MacFarlene, J. 2004. "In What Sense (If Any) Is Logic Normative for Thought?" Unpublished manuscript. Presented at the Conference on the Normativity of Logic organized by Central Division APA 2004.

Harman, G. 1986. *Change in View: Principles of Reasoning*. Cambridge: MIT Press.

Hayek, A. 2001. "Probability, Logic, and Probability Logic." In L. Goble (ed.). *The Blackwell Guide to Philosophical Logic*. Oxford: Blackwell.

Hitchcock, D. 2017. *On Reasoning and Argument, Essays in Informal Logic and on Critical Thinking*. Springer.

Shapiro, S. 2014. *Varieties of Logic*. Oxford: Oxford University Press.

Steinberger, F. Forthcoming. "Three ways in which logic might be normative." *Journal of Philosophy*. (http://floriansteinberger.weebly.com/uploads/5/7/9/5/57957573/three_ways_in_which_logic_m ight_ be_normative.pdf)

van Eemeren, F. H. and Grootendorst, R. 2004. *A Systematic Theory of Argumentation. The pragma-dialectic approach*. Cambridge: Cambridge University Press.

Varga, A., Stenning, K., and Mortignon, L. 2015. "There Is No One Logic to Model Human Reasoning: the Case from Interpretation." In U. Furbach and C. Schon (eds.). *Proceedings of the first workshop on bridging the gap between human and automated reasoning*. Berlin: 32–46.

Tarski, A. 1956. *Logic, Semantics, Metamathematics*. Oxford: Clarendon Press.

Croatian Journal of Philosophy
Vol. XVIII, No. 54, 2018

Some Limitations on the Applications of Propositional Logic

EDI PAVLOVIĆ
University of Helsinki, Helsinki, Finland

This paper introduces a logic game which can be used to demonstrate the working of Boolean connectives. The simplicity of the system turns out to lead to some interesting meta-theoretical properties, which themselves carry a philosophical import. After introducing the system, we demonstrate an interesting feature of it—that it, while being an accurate model of propositional logic Booleans, does not contain any tautologies nor contradictions. This result allows us to make explicit a limitation of application of propositional logic to those sentences with relatively stable truth values.

Keywords: Logic gate, logic game, a priori, propositional logic.

1. *Introduction*

In this paper, we imagine first introducing (propositional) logic to a rational thinker via a simple logic game. It will be obvious this is a perfectly adequate way of explaining the Boolean connectives. At the same time, however, the bare-bones simplicity of the system will turn out to lead to some interesting meta theoretical properties, which themselves carry a philosophical import. That philosophical import has to do with the truths of logic, and the high rank these occupy among things we can know *a priori* (cf. Boghossian 2003, Harman 1996, Kitcher 1980, Peacocke 2005). It is useful to think that "[...] in general a logical truth is a statement which is true and remains true under all reinterpretations of its components other than the logical particles." (Quine 1961: 22–23). Sticking to propositional logic and the paradigmatic case of the law of excluded middle, $P \vee \neg P$, this idea would mean that this proposition's truth depends only on the logical symbols '$\vee$' and '$\neg$' or, in other words, is independent of what *P* might be (as long as it is a proposition). The properties of this system cast a doubt on that notion. In the following section of this paper, we will present the system in question, and its connection with propositional logic. In the central section, we demonstrate its meta-theoretical properties. Finally, the closing section of this paper takes a slightly bigger-picture perspective with an examination of the philosophical implications of the system.

2. *Lock-Key* game

The system at hand is presented in a form of a logic puzzle, one that asks whether a certain key opens a certain lock. Drawing inspiration from a computer infrastructure, locks are built of elements closely resembling logic gates, but with symbols altered to resemble the standard notation of Booleans in logic, including a specific orientation of the notation.

2.1 *Lock* elements

Lock consist of three kinds of elements, each with two subtypes: circles, squares and triangles, determined by the number of connections they have to other elements (one, two and three, respectively).

> *Definition 1* (*Circle*) Circle elements are the input/output elements of the lock. A number of *Input* circles are located at the bottom (i.e. "beginning") of the lock, and a single *Output* circle is located at the top (i.e. "end") of the lock. Note that a lock by definition contains only one output (and does in fact contain one).

Graphically we represent these elements as:

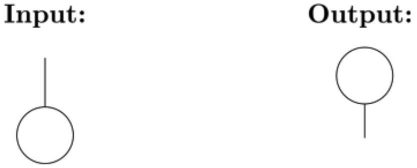


Fig. 1: Circle elements

> *Definition 2* (*Square*) Square elements have two connections—the one below considered its input and the one above considered its output. The types of this kind of an element are labeled *Same* and *Other*.

Graphically we represent these elements as:

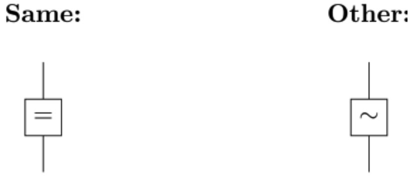


Fig. 2: Square elements

Definition 3 (*Triangle*) Triangle elements have three connections—the two below are considered its input and the one above considered its output. The types of this kind of an element are labeled *Minimum* and *Maximum*.

Graphically we represent these elements as:

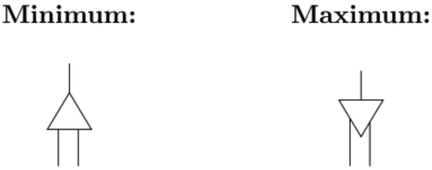


Fig. 3: Triangle elements

Note that Squares and Triangles are all *functions*, i.e. for any input they have only one output. The purpose of the Square element *Same* is purely for the sake of legibility, and will be discussed in the next section. The reader is probably aware of the purpose of the elements *Other*, *Minimum* and *Maximum*, but let us just note here that the shape of the latter two were chosen to facilitate memorization—*Minimum* is the widest at the bottom of the element, while *Maximum* is widest at the top.

We now proceed to introduce another element of the game—the keys, which serve as the primary input for the whole structure of a lock.

2.2 *Key* elements

Definition 4 (*Key*) A key is an ordered n-tuple $\langle v_1 \dots v_n \rangle$ where $v_i \in \{0,1\}$. A key is *a key to a lock* $\boldsymbol{\mathcal{L}}$ just in case n is the number of input elements of the lock $\boldsymbol{\mathcal{L}}$.

After introducing the elements, we now put them all together to describe how the “game” is played.

2.3 *Building a lock*

> *Definition 5* (*Lock building procedure*) A lock is built bottom up, and starts with a number of *Input* elements. We then add the other elements, with each of their inputs coming from an output of a lower element, and finally we add an *Output* element. The elements that some element ε gets its input from are called the *immediate predecessors* of ε.

Two limitations are that the output of each Square and Triangle is an input of some other element and that there is only one *Output* element.

In case we wish to consider only certain sections of a lock, we refer to them as *sublocks*:

> *Definition 6* (*Sublock*) Any element is a part of the same sublock as itself. If an element is part of a sublock, then the element(s) it gets its input from are also parts of the same sublock. We refer to sublocks as sublocks of their topmost element.

Example 7 An example of a correctly built lock is the following:

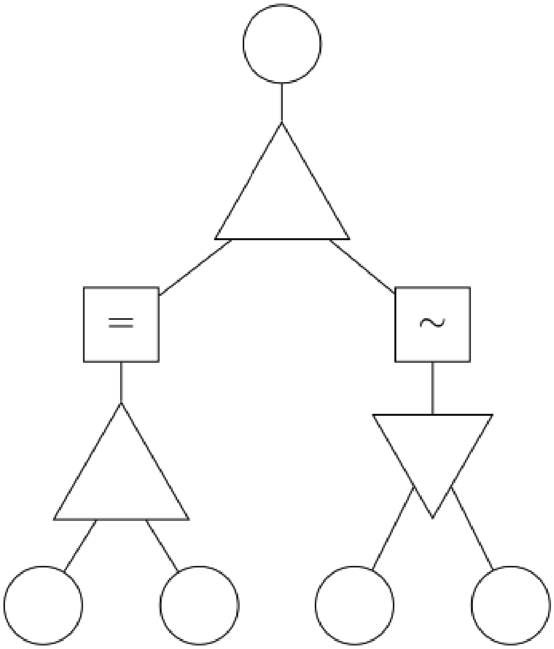


Fig. 4: An example of a lock

2.4 *Semantics of the game*

After introducing the lock and key elements, we now proceed to define their interactions:

> *Definition 8* (*Semantics*) The elements of a lock $\mathcal{L}$ produce the values as follows. Going from left to right, the *Input* element i takes the value v_i from the key to the lock $\mathcal{L}$. The element *Same* produces as its output the same value as it input, element *Other* produces the other value than its input, *Maximum* produces the greater of the two values in its input, and *Minimum* the lower. Finally, if the *Output* element takes the value 1 we take the lock to be "opened," and "closed" otherwise.

Example 9 Going back to the example lock from the Example 7, the only key to produce an "open" lock would be $\langle 1,1,0,0 \rangle$. It is left as an exercise to the reader to verify this.

2.4.1 *Connection to propositional logic*

It should be clear to the reader that

> *Observation 10* The semantics of the elements *Other*, *Maximum* and *Minimum* are precisely those of the Booleans *Not*, *Or* and *And*, respectively.

As an illustration, the Example 9 above would correspond to the formula $(A \wedge B) \wedge \neg(C \vee D)$ with the main connective being the conjunction, the left conjunct being a further conjunction, and the right one a negation of a disjunction. It is again left as an exercise to the reader to verify this formula is assigned 1 only when the variables *A, B, C* and *D* are assigned the values 1, 1, 0, 0, respectively.

Given this, it should be clear how this game will allow our thinker to understand propositional logic. We can also see why the element *Same* might be useful, as it reflects the idea that a subformula (which itself corresponds to a sublock) is assigned a value, but that value gets picked up by some other connective at a later stage. One can further utilize the game to introduce the semantic tables by considering how many possible keys there are for each lock. Note also that the shape of the triangle elements, while making sense in the context of *Minimum* and *Maximum*, also make it easier to distinguish the common disjunction and conjunction symbols.

We will now proceed to consider some metatheoretical properties of the game.

3. *Metatheory of the game*

As we have seen in the Example 7, all but one possible key produced a closed lock. The question that now arises is whether we can take this one step further by building a lock that is "unopenable."[1] Note that:

> *Observation 11* This question is equivalent to asking whether we can have a lock which all keys open, since we can transform one into the other by adding one *Other* element just before the *Output* element.

To encompass both of these, we will use the expression "forced."

> *Definition 12* (*Forcing a lock*) A lock can be *forced* if it can be made in such a way as to produce the same output for any possible key.

The answer to the former question is simply 'No' and the demonstration of this fact will be the central result of this paper.

> *Theorem 13* No lock in the *Lock-Key* game can be forced.

Proof. By showing no sublock can be forced, by induction on the last element.

Basic case. Input elements cannot be forced, by Definition 8.

Inductive case. If the last element is *Same*, its sublock can only be forced if the sublock of its immediate predecessor is forced. If the last element is *Other*, its sublock can likewise only be forced if the sublock of its immediate predecessor is forced, given Observation 11. Similarly for the Triangle elements, we can only force them if we can force the sublocks of their immediate predecessors.

Given Observation 10, it is clear that an unopenable lock would correspond to a contradiction, and a lock that every key opens to tautology.

In light of this, an interesting corollary follows:

> *Corollary 14 Lock-Key* system contains no tautologies and no contradictions.

4. *Philosophical import*

What, if any, is the philosophical lesson and importance of this result? To see this, let us consider how to reconcile the apparent strain that exists between Observation 10 and Corollary 14. After all, the former tells us the elements of the game correspond to the Booleans, but those, in opposition to the corollary, do produce tautologies and contradictions (given our choice of the Booleans, the examples we'll focus on are $P \vee \neg P$ and $P \wedge \neg P$). The result that resolves the discrepancy, and is the main, if modest (given the limited scope of this paper) philosophical claim of this paper is the following:

> *Theorem 15* It is not the case that the tautologies [contradictions] of propositional logic are true [false] purely in virtue of the Booleans occurring in them.

[1] The author owes gratitude to Dragana Sekulić for bringing this question to their attention.

Proof. To see this, one should only note that the locks also contain *Input* elements, which correspond to the propositional variables. However, the semantics of these elements are not exactly alike—recalling the basic step of the proof of Theorem 13, *Input* elements cannot be forced. But the propositional variables can—the second occurrence of the variable P in either the example tautology or contradiction can have only one value—that of the first occurrence. This difference reveals an additional ingredient one needs to "cook up" a sentence of propositional logic that is always true or false—a certain type of behavior of the propositional variable(s).

Note that this argument does not show, nor does it in fact intend to do so, that the law of excluded middle (or the law of non-contradiction) is not true. Therefore, it is not a version of a "deviant logician" argument (cf. Williamson 2006, Sullivan 2014). The *Lock-Key* system corresponds to a system without a very basic, and by and large implicit, feature of propositional logic, that the propositional variables do not alter their value within an assignment. Such a system could be, e.g. one where we use a propositional variable only once in a formula. Alternatively, it could be one with such assignments which would allow the change of value while a formula is in use. This isn't entirely far-fetched, e.g. "I never wrote this sentence before." Rather, it just illustrates that logic is more regimented than we normally notice.

In explaining logic to our thinker using this game, we will have to explicitly and separately introduce limitations on the truth values of propositional variables. The very fact that the system is so basic allows it to make this feature of propositional logic apparent.

References

Boghossian, P. A. 2003. "Epistemic analyticity: A defense." *Grazer Philosophische Studien* 66 (1): 15–35.

Harman, G. 1996. "Analyticity regained?" *Noûs* 30 (3): 392–400.

Kitcher, P. 1980. "A priori knowledge." *The Philosophical Review* 89 (1): 3–23.

Peacocke, C. 2005. "The a priori." In F. Jackson and M. Smith (eds.). *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press.

Sullivan, A. 2014. "Logical deviance and the constitutive a priori." *Discusiones Filosóficas* 15: 67–85.

Quine, W. V. O. 1961. "Two Dogmas of Empiricism." In his *From a Logical Point of View*, 2nd ed. Harvard University Press.

Williamson, T. 2006. "Conceptual truth." *Aristotelian Society Supplementary Volume* 80 (1): 1–41.

Croatian Journal of Philosophy
Vol. XVIII, No. 54, 2018

How Gruesome are the No-free-lunch Theorems for Machine Learning?

DAVOR LAUC
University of Zagreb, Zagreb, Croatia

No-free-lunch theorems are important theoretical result in the fields of machine learning and artificial intelligence. Researchers in this fields often claim that the theorems are based on Hume's argument about induction and represent a formalisation of the argument. This paper argues that this is erroneous but that the theorems correspond to and formalise Goodman's new riddle of induction. To demonstrate the correspondence among the theorems and Goodman's argument, a formalisation of the latter in the spirit of the former is sketched.

Keywords: Induction, the problem of; No-free-lunch theorems, The new riddle of induction.

1. *Introduction*

Right from its beginning, the development of artificial intelligence and machine learning prompted many interesting philosophical debates and provoked some interesting philosophical questions. On the flip side, researchers in these fields often encounter or rediscover classical philosophical problems. One such case is the identification of the no-free-lunch (NFL) theorems—the famous negative results in machine learning—as a reiteration or even a formalisation of the Hume's problem of induction. The distinguished machine-learning researchers like Christophe Giraud-Carrier and Pedro Domingos state, respectively:

> It then becomes apparent that the NFL theorem in essence simply restates Hume's famous conclusion about induction having no rational basis... (Giraud-Carrier 2005)

and:

> ...This observation was first made (in somewhat different form) by the philosopher David Hume over 200 years ago, but even today many mistakes in machine learning stem from failing to appreciate it. (Domingos 2012)

Even the originator of the first form of the NFL theorems, David Wolpert, fifteen years after he proved the theorem, joins the information cascade and claims that:

> …these original theorems can be viewed as a formalisation and elaboration of concerns about the legitimacy of inductive inference, concerns that date back to David Hume… (Wolpert 2013)

This paper argues that the NFL theorems, although vaguely connected to the classical philosophical problem of induction, do not restate the Hume's problem, but rather the associated Nelson Goodman's argument.[1] We claim that the NFL theorems are closely related to Goodman's new riddle of induction (NRI), to the extent that they are one possible formalisation of the riddle. Additionally, we would like to pose the question of the relevance of the NFL to the vast philosophical discussion on NRI, as the relationship is yet to be researched. The related, reversed, the issue is the relevance of NRI to NFL and the question as to whether the machine-learning community could benefit from the almost 70 years of fruitful discussion about Goodman's argument.

1.1 No-Free-Lunch Theorems

The first form of NFL theorem was proven by Wolpert and Macready, 1992, in the context of computational complexity and optimisation research (Wolpert and Macready 1995; Wolpert 1992). He later proved the variant of the theorem for the supervised machine learning (Wolpert 1996). For the sake of our argument, we will sketch the proof of the simplified version of the theorem for supervised learning based on the work of Cullen Schaffer (Schaffer 1994).

In the simplest, discrete settings of the machine learning of a Boolean function, training data X consists of the set of binary vectors representing a set of attributes that are true or false for each instance of the binary function—concept. Each vector is labelled as a positive or negative example of a concept we want to learn. The machine-learning algorithm L tries to learn a target binary function y; a true concept from this set of examples. Training dataset is always finite with some length n, and the relative frequency of data feed to L is defined by probability distribution D. In a context more familiar to the philosophers, this problem of machine learning can be seen as a guessing a true form of a large n-ary truth function from the partial truth-table, where most of the rows are not visible.

The key performance indicator of a machine learner is a generalisation performance, with the accuracy of the learner found within the data outside the training dataset. Modern machine-learning algorithms can easily "memorise" data from the training dataset, and perform poorly on the "unseen" data, leading to the problem known as overfitting. So,

[1] We suppose that the new riddle is a different issue from the classical problem of induction, what is the received position with a few notable exceptions like (Magnus 2006).

the success of the learner is measured by how well it will generalise, and how well it performs on the novel data. In the simple setting of binary concept learning, the baseline of the generalisation accuracy of a learner, *GP*(*L*) is the random guess, with the accuracy of the novel data being 0.5. This is the performance we will expect on average if we use the toss of a coin to decide, for an unseen example, whether it belongs to our target concept or not. Clearly, we want any learner to perform better than this.

The NFL theorem claims that, for any learner *L*, given any distribution *D* and any *n of X*

$$\frac{\sum_{f \in Y} GP(L)}{|Y|} = 0.5$$

where *Y* is a set of all target functions, all possible concepts that can be learned.

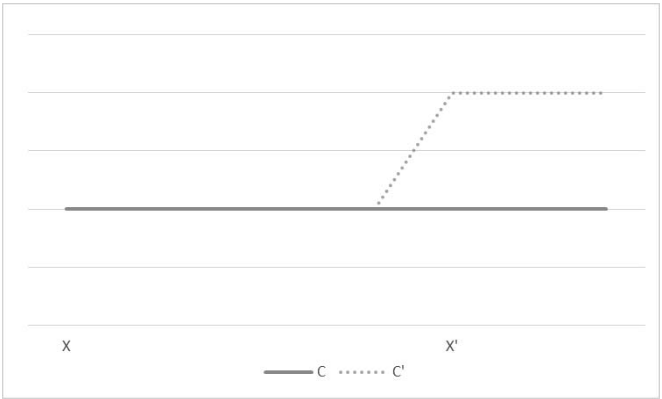
So, the theorem states that, on average, the generalisation performance of any learner is no better than random guessing. All learning methods, from the simple decision trees to the state-of-the-art deep neural networks, will perform equally when all possible concepts are considered.

This result, unanticipated at least on the first sights, does bear some resemblance to the discrepancy between the results of the argument and our expectations in the case of the Hume's argument. However, it does not claim that we cannot learn anything from the training data or experience, but that we can learn everything, which is, arguably, the point of Goodman's argument. The resemblance to the NRI will be more evident from the sketch of the proof of the NFL theorem. The basic idea is very simple: for any concept that the learner gets right, there is a concept that it gets wrong or, in Goodman's lingo, for every "green" concept there is a "grue" concept. The "grue" concept is constructed similar to the NRI argument, in that it agrees on all observed data—data in the training dataset—with the "green" concept, and it is bent on all non-observed data.

More formally, for every concept *C* that *L* learns to classify well—say it classifies *m* novel examples accurately—there is a concept *C′* that *L* learns where all *m* examples will be misclassified. *C′* is constructed as follows:

$$C' = \begin{cases} C \ if \ x \in X \\ \neg C \ if \ x \notin X \end{cases}$$

Visually, this simple construction of the concept *C′* corresponds to the Wittgenstein-Goodman "bent predicate" (Blackburm 1984), where X represents observed data (training dataset) and X' unobserved data.



From the perspective of the main measure of the success of the learning—generalisation accuracy, for every accuracy improvement *a* over the baseline for a concept *C*, there is a concept *C′* that will offset the improvement of the accuracy by –*a*. Consequently, the improvement in accuracy for any learner over all possible concepts is zero. It is possible to generalise this result to the more general learning settings, and many extensions of the theorem are proven (Joyce and Herrmann 2018; Igel and Toussaint 2005).

1.2 *Is the No-free-lunch Theorem the New Riddle of Induction?*

Although NFL bears a strong resemblance to NRI, it seems worthy to analyse differences and similarities between these two results. Let's start with the similarities. Both arguments are about inductive inference, about inferring from the known to unknown, and from the observed to the unobserved data. Both arguments imply that there are too many inductive inferences that can be inferred. Furthermore, both arguments seem to draw empirically inadequate conclusions, contrary to scientific practice and common sense expectation. Nobody is expected to conclude that all emeralds are grue, and neither that random guess is an inductive strategy as good as any other.

The key resemblance is in the construction of the arguments, the split of the evidence and the bend in the unobserved data. In most of the NRI arguments, we split the evidence into observed and unobserved (sometimes to some point in the future). Equally, in the NFL, data are split into observed, training dataset and the unobserved data to which the learning algorithm should generalise. In both arguments, the other counter-concept, *grue* or *C′*, is constructed in the same manner. It agrees on the observed data and bends on the unobserved data.

Regarding the differences, firstly, there is a difference in the argument contexts. NRI was made in the philosophical, theoretical context of the logic of confirmation and pragmatic vindication of induction, while the NFL was made in the technological context of artificial intelligence and computing. The aim of the arguments also differs, at least at first glance. The intention of NRI, at least in Goodman's initial form (Goodman 1946; Goodman 1983), was to recognise one of the problems in the logic of confirmation—the demarcation between projectable and non-projectable predicates. On the other hand, the objective of the NFL was to demonstrate that there is no single best algorithm, initially for the optimisation and search, and later for supervised learning.

The biggest difference seems to be in the scope of quantification. The no-free-lunch theorem quantifies over all learners and all concepts, while Goodman's argument seems to be about constructing one particular example. However, NRI can be reformulated to have a similar quantificational structure as the NFL.

2. *The New Riddle of Induction as the No-free-lunch Theorem*

Goodman's argument, at least in one of its interpretations, can be rephrased in the NFL fashion as follows. Let's define *P*, the degree of projectability, as the generalisation accuracy in the settings of learning Boolean function. Let's define *L* to be a language, understood informally as a frame of reference or level of abstraction (Floridi 2008). It is also possible to define language using a more formal framework like a web ontology (McGuinness 2004) or the formal concept analysis (Ganter 2012). Finally, let's define *I* as an inductive strategy or, as in Goodman's original argument the logic of confirmation. Then, we can state the new riddle as:

$$\forall I\ \forall L \frac{\sum_{x \in L} P(x)}{|L|} = 0.5$$

Claiming that, for any inductive strategy, the degree of projectability over all possible languages is 0.5, what is zero improvement over a random guess. The formal condition, by the NFL condition, is that the languages are unrestricted or that they are closed under permutation (Schumacher 2001). The proof of such stated NRI would be the same as the proof for NFL—for every concept *C* with a degree of projectability *p*, there is a concept *C′* with the degree of –*p*, and for every "green" concept there is a "grue" concept.

3. *Discussion and Future Work*

The takeaway of this formalisation would be one of the lessons that Goodman has taught us—the importance of the language for the induction, or the impossibility of empirical investigation without some predefined language that we bring to the process. It is interesting to compare this with the conclusion that the same researcher from the machine-learning community draws from the NFL—there is no learning without bias, there is no learning without knowledge (Domingos 2015).

If we accept the above formalisation as one of the possible interpretations of the NRI argument, it would be interesting to examine how the NFL can contribute to the NRI exploration. One of the approaches is to investigate how different NFL conditions apply to NRI. For example, if we restrict a set of the concepts to some subset of all those possible, it has to be closed under permutation (C.U.P.) for NRI to hold (Schumacher 2001), and the fraction of such subsets is tiny. Another potential route of exploration is to research the relevance of non-uniform distribution of target functions/concepts for arguing that NRI works (Igel and Toussaint 2005) and results that the NRI does not extend to the continuous domains (Auger 2007).

On the flip side, it would also be interesting to explore possible "solutions" to the NFL theorem from the NRI perspective. Is it possible to use Goodman's pragmatic solution in limiting the NFL by formalising his concept of entrenchment? It would also be interesting to research additional constraints on NFL languages using Davidson's approach to NRI (Davidson 1966). Finally, one of the biggest social and ethical problems of machine learning, especially deep learning, is the problem of the interpretability of models. It would be interesting to research whether the philosophical explorations in the NFL could help in this direction.

References

Auger, A. a. 2007. "Continuous lunches are free!" *Proceedings of the 9th annual conference on Genetic and evolutionary computation.* ACM.

Davidson, D. 1966. "Emeroses by other names." *The Journal of Philosophy* 63 (24): 778–780.

Domingos, P. 2012. "A few useful things to know about machine learning." *Communications of the ACM* 55 (10): 78–87.

Domingos, P. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World.* Hachette UK.

Floridi, L. 2008. "The method of levels of abstraction." *Minds and machines* 18 (3): 303–329.

Ganter, B. a. 2012. *Formal concept analysis: mathematical foundations.* Springer Science.

Giraud-Carrier, C. a. 2005. "Toward a justification of meta-learning: Is the no free lunch theorem a show-stopper." *Proceedings of the ICML-2005 Workshop on Meta-learning.*

Goodman, N. 1946. "A Query on Confirmation." *The Journal of Philosophy* 43 (14): 383–385.

Goodman, N. 1983. *Fact, fiction, and forecast.* Cambridge: Harvard University Press.

Igel, C. and Toussaint, M. 2005. "A no-free-lunch theorem for non-uniform distributions of target functions." *Journal of Mathematical Modelling and Algorithms* 3 (4): 313–322.

Joyce, T. and Herrmann, J. 2018. "A Review of No Free Lunch Theorems, and Their Implications for Metaheuristic Optimisation." *Studies in Computational Intelligence* 744: 27–51.

Magnus, P. D. 2006. "What's new about the New Induction?" *Synthese* 148 (2): 295–301.

McGuinness, D. L. 2004. "OWL web ontology language overview." *W3C recommendation* 10 (10).

Schaffer, C. 1994. "A conservation law for generalization performance." *Machine Learning Proceedings* 1994: 259–265.

Schumacher, C. M. 2001. "The No Free Lunch and Description Length." In E. G.-M. L. Spector. *Genetic and Evolutionary Computation Conference (GECCO 2001).* San Franisco: 565–570.

Wolpert, D. 1992. "Stacked generalization." *Neural networks* 5: 241–259.

Wolpert, D. 1996. "The Lack of A Priori Distinctions between Learning Algorithms." *Neural Computation* 1341–1390.

Wolpert, D. H. 2013. "Ubiquity symposium: Evolutionary computation and the processes of life: What the no free lunch theorems really mean: How to improve search algorithms." *Ubiquity, December 2013.*

Wolpert, D. and Macready, W. 1995. "No Free Lunch Theorems for Search." *Technical Report SFI-TR-95-02-010 (Santa Fe Institute).*

Croatian Journal of Philosophy
Vol. XVIII, No. 54, 2018

Book Discussion

Reconciling Poetry and Philosophy: Evaluating Maximilian De Gaynesford's Proposal

IRIS VIDMAR and MARTINA BLEČIĆ
University of Rijeka, Rijeka, Croatia

Poetry and philosophy have had a long and convoluted relation, characterized often by mutual antipathy and rarely by mutual acknowledgment and respect. Plato was one influential philosopher who trashed poetry's capacities to trade in the domain of truth and knowledge, but it was J. L. Austin who blew the final whistle by dismissing it as non-serious. And while for many poets that was an invitation to dismiss Austin, for many philosophers that was a confirmation of the overall discomfort they had already felt with respect to poetry. Just how wrong both parties were in this standoff is revealed in the latest book by Maximilian De Gaynesford, The Rift in the Lute: Attuning Poetry and Philosophy, which calls for a dismissal of the separation of the two and for their mutual cooperation. In this paper, we look at De Gaynesford's proposal, mostly praising its strong points and occasionally raising doubts regarding its success.

Keywords: J. L. Austin, philosophy, poetry, Maximilian De Gaynesford

It is hardly an exaggeration to say that philosophers have, for the most part, ignored poetry (see Ribeiro 2009; Gibson 2015). Luckily, things are changing and poetry has started to attract attention. The latest book by Maximilian De Gaynesford, *The Rift in the Lute: Attuning Poetry and Philosophy,* is a much welcome addition to this trend, one which will for sure initiate its own wave of responses. Gaynesford does not aim to say much about the aesthetic or artistic value of poetry, and he doesn't dwell on issues of definition. Rather, he deals with one of the most influential claims regarding poetry ever made: J. L. Austin's views on poetry as 'not serious'. Determined to prove Austin wrong, Gaynesford sets out to develop a new account of poetry and to suggest new ways in which to view the convoluted relationship between poetry and philosophy.

As Gaynesford argues, it is of particular importance for analytic philosophy to turn to poetry, and to do so from the perspective of a speech act theory: it is here that "relations between literature and philosophy are at their worst" and "their antipathy" at its deepest (12-3). Consequently, to reconcile them, that is the place to start. How? By following the project of attunement—"a mutually shaping approach in which we really do philosophy in really appreciating poetry, doing the literary criticism necessary for this" (9). Gaynesford's project is thus a matter not of applying philosophy to poetry—thus doing a philosophically minded literary criticism—but the one of "exercising our critical engagement with poems *in* engaging with philosophy, and exercising our critical engagement with philosophy *in* engaging with poems." (10). Such joint collaboration is envisioned as a win-win situation for both: "The opportunity to appreciate philosophical distinctions and discriminations in poetry can improve our ability to discriminate features of philosophical significance. And this opportunity to grapple anew with philosophy in turn heightens our capacity to appreciate what is rich and subtle in poetry, which returns us more richly provided to pursue philosophy, from where we can go back more generously supplied to appreciate poetry, and so on, back and forth" (11).

At the centre of the attunement project is a radical turn from the way philosophers (and critics, to some extent) usually approach poetry, namely from the point of view of its alleged disconnection from the truth. While philosophers mostly attend to poetry in order to either show or to dispute that poetic language is incommensurable to the epistemic goals of conveying truth,[1] Gaynesford sets his theory in a completely different setting: that of philosophy of action. Rather than approaching poetry as a set of true or false statements or descriptions, Gaynesford suggests that saying things in poetry—uttering poetically—is not a matter of stating things but of doing things. Those familiar with Gaynesford's philosophical profile will not be surprised to learn that his account is motivated by J. L. Austin's famous statement about poetry not being serious.

These brief introductory remarks suffice to position Gaynesford's book within the relevant theoretical framework: in the first part, Gaynesford works out the details of his attunement project by carefully and informatively elaborating on the ways in which Austin dismisses poetry as serious (ch. 1), and by examining how poets and critics (ch. 2), as opposed to philosophers (ch. 3), reacted to Austin's remarks. He then moves on (chs. 4 and 5) to show how these debates reflected on poetry's connection to truth, and ends by arguing for a paradigm shift (chs. 6

[1] See in particular essays gathered in Gibson (2015).

and 7): poetry should be viewed as a form of action, and poetic utterances as utterances which actually do things. Once this approach is taken, a need is generated to account for the responsibility and commitment of those creating poetry. In the second part, Gaynesford first analyzes (ch.7) what he calls 'the Chaucer type utterances' and, having explicated their main features (chs. 9-11), applies his account to numerous Shakespeare's sonnets. It is in this part that he engages in a rather insightful form of literary criticism, one which presupposes an attuned relationship between poetry and philosophy, showing the drawbacks of those critical views on Shakespeare which failed to appreciate what a philosophically minded reader can see in the sonnets, and what the sonnets can reveal to the reader open to philosophical concerns.

The richness of Gaynesford's theoretical framework does not imply lack of detailed and meticulous exploration of its constituents, including, in the opening and closing chapter, a detailed analysis of real world examples in which poetry was taken seriously enough for its creators to face serious legal issues. In many ways, his interpretation of the relationship between poetry and philosophy is insightful, primarily due to his exhausting overview of various poets, critics and philosophers who had something to say on the topic. Gaynesford's analysis along these lines will challenge the somewhat dominant view according to which philosophers, on the whole, are hostile to poetry, and according to which poetry has, for the most part, been the "victim of Austin's efficiencies". It will also cast doubt on the way Austin's views on poetry are most commonly interpreted. As Gaynesford argues, though Austin represents poetry as non-serious use of language, where language is not used in the normal way, or is used in hollow and void way, parasitic upon the normal use (39), he neither argues for these claims, nor does he clarify their meaning. Austin's crucial failure is the fact that "the combination of high-handedness and half-heartedness" in his writings on poetry, as well as the examples he chose to support his view, "give the strong impression that he recognized something forced about ... this insistence that poetic utterances are *not* to be understood in terms of things that are done" (259, emphasis original). In other words, Austin's remarks "make no distinction between types and instances of poetic utterances", offer "no arguments to demonstrate that *no* poetry is serious", and ignore the ambiguity of notions he uses to express the alleged non-seriousness (39). The dominantly poetic manner itself, in which Austin writes about poetry, as compared to his other writings, reveals, on Gaynesford's reading, that Austin himself has hard time accepting what he says—his argument, in other words, "resists taking itself seriously" (44).

Gaynesford further argues that most of the poets who set out to respond to Austin failed to properly engage with his views, mostly due to a prejudice they harboured about philosophy's overall distrust of philosophy. And while critics have for the most part turned Austin into a bad guy unappreciative of the value of poetry, Gaynesford argues that Austin is far more appreciative of poetry than Plato or Frege ever were; his bad reputation is a consequence of critics' failure to engage properly with philosophical views. The critics are, generally, just as "careless and disdainful" (58) towards Austin as Austin is toward poetry. Sadly, philosophers are no better. In failing to properly engage with Austin's remarks, they "expose their own prejudice against poetry: they condone the insults, neglect the tensions and contradictions, hide the ambiguities, and assume a determinacy where all is vagueness". Consequently, "no wonder so much that is philosophically significant in poetry is ignored, and so much in philosophy that is relevant to the appreciation of poetry goes unrecognized" (68-9). As Gaynesford further demonstrates in the fifth chapter, another failure on the part of philosophers relates to the fact that for the most part, they analyzed poetry as if poetry was to be evaluated from the perspective of whether or not it told the truth about the world. Such misconception is itself an outcome of the 'governing assumption' among philosophers, one which Austin himself set out to refute, according to which it is the main function of language to describe things. Whereas Austin wanted to show that language also does things, i.e. that we do things when we utter propositions, philosophers remained focused on analyzing poetry' s success or failure to correctly describe things, and completely ignored the fact that it too can get things done. On Gaynesford's view, "this way of approaching poetry renders essential features of poetry invisible and distorts literary criticism" (261). To amend such mistreatment, Gaynesford offers his own, attuned account.

Gaynesford's analysis showed that, appearances aside, philosophers and poets do agree that poetic uses of language are exempt from issues of commitment and responsibility. However, it is precisely this presupposition that is wrong, which can only be acknowledged once poetry is approached from the standpoint of philosophy of action, and within it, from the perspective of a speech act theory. Within such "realigned debate", issues of commitment and responsibility can be reassessed. As Gaynesford argues, poets can use language seriously, for "to be serious is to acknowledge what is required if one is to be taken seriously: a commitment to be reasonably clear about what one means, to be willing to explain what one says, to account for what one claims. And it is not only possible but actual that poets commit themselves responsibly in these various ways (for example, in essays, reviews, manifestos, interviews)" (110). Crucial questions that are to be asked with respect to poetry under such an account are questions "about who is accountable for a particular utterance, what was intended by some particular choice of words, whether the action performed is one for which its author can be praised or blamed"—questions, as Gaynesford argues, that already "define literary criticism and which commentators on poetry have placed at the centre of their endeavours" (112). As a crucial example of a poet who used poetry in this manner, Gaynesford refers to Chaucer, whose poems are riddled with what Gaynesford calls Chaucer-type utterances. These utterances are composed of a first person concatenated with a verb in the present indicative active (i.e. I dedicate, I direct) and they correspond to what would in non-poetry be equivalent to 'explicit performatives'—though naturally the proper classification is complicated by the fact that such performatives are further divided into various subgroups. As Gaynesford warns us, there is a considerable disagreement regarding this type of utterances, but he nevertheless goes on to elaborate on four main features they exhibit: doing (in uttering the relevant sentence, the speaker does something beyond uttering), phrasing (the sentence uttered contains a sentential clause consisting of a subject term (the first person pronoun in the nominative) concatenated with a verb of doing (first-person singular, resent tense, indicative mood, active voice) combined with an explicit or implicit 'hereby' or its equivalent); naming (the verb in the sentential clause is a word for what the speaker does in uttering the sentence) and securing (the act named by the verb in the sentential clause is assuredly performed in uttering the sentence). Given that these four features can be employed in variety of ways, analysing various combinations in which they come together in any given poem offers additional chance for philosophers of language to analyse them, but it also offers to critics a possibility to analyze such poems from different perspectives—after all, that is what the attunement approach is meant to initiate.

To support his claims, in the final chapters of the book Gaynesford turns to Shakespeare's sonnets and analyses how the great bard uses the four features of Chaucer-type utterance. "Recognizing the dramatic salience of the type has the power to develop and change the way we see the sequence [of sonnets] as a whole, as well as the individual poems of which it is composed" states Gaynesford (263), and goes on to show numerous ways in which Shakespeare deploys the four features, often modifying them, even to the point where it is not altogether certain whether the Chaucer type has in fact been used. However, such ambiguity is identified as the source of variations of meaning of the sonnets, which result from different ways in which phases and lines in poems might be understood. Such an approach enables Gaynesford to, among other things, analyse ways in which some of Shakespeare's sonnets are imbued with Cartesian type of scepticism, with considerations regarding limits and limitations, obligations and duties, one's solipsistic worries, etc. It further enables him to analyse ways in which different poems reflect on poetry as a mode of using language and on poetry as a form of action, which in turn draws attention to the means we have at our disposal to study poetry, and to the philosophical issues generated by these means.

Gaynesford's account of poetry departs from some of the ways in which poetry is traditionally analyzed, which will either seem like a welcome new paradigm to be happily embraced, or like a dead-end street to be quickly abandoned. There are, we think, many important aspects of his proposal which give us a more profound understanding of poetry, and his ambition to reconcile poetry and philosophy seems promising—though the question remains whether those untrained in philosophy could appreciate poetry generally and individual poems in ways Gaynesford envisions. Gaynesford does not place much emphasis on the aesthetic aspects of poetry, and when he does, he subjects them to the goals that poetic utterances containing such aesthetic properties are to realize. By thus instrumentalizing what for many is the crucial aspect of poetry and poetic experience, Gaynesford's theory might be dismissed by authors who oppose subjecting poetry to philosophical concerns. On the other hand, there have been attempts recently, predominantly made by literary scholars or poets themselves, to show ways in which poetry (and literature more generally) manages to bring about some more tangible changes, whether in the mindset of individual reader or within wider social groups and cultures.[2] For those who appreciate such approach to poetry and who share such views on its potential, Gaynesford's book might serve as an insightful pointer on how poetry might have such power, as his account is well suited to explain the tendency of critics to talk of poetry as achieving (or having the effect of initiating) intellectual paradigm changes. Another way in which Gaynesford's account is inspiring relates to what it might add to our understanding of the poetic language and everyday communication. Gaynesford notes that "examples of poetic utterances reveal underlying distinctions in the way poetry does things with words. The addition of new categories of actions, some peculiar to poetry, reveals ways in which philosophy can increase knowledge of language-use by attending to poetry." (114) It would be theoretically useful to identify those speech-acts 'peculiar to poetry' and see in which relation they stand with other kinds of speech acts. If regular speech acts as stating or promising can be incorporated in poems, is there a place for poetic speech acts in everyday communication? Do poetic speech acts turn everyday communication into poetry or is it so that they cannot be part of it since it would mean that they are not exclusively poetic? Identifying those speech acts 'peculiar to poetry' would be a good start in analyzing the relation between poetry and other uses of language.

The backbone to Gaynesford's proposal is the idea of poetry as a form of action. To many, it is this particular premise in his overall account that might be the hardest to swallow. Of course Gaynesford is aware of that, and he dedicates the entire chapter 5 to smooth some possible objections. Three he sees as the most pressing: first, whether poetic utterances can indeed be understood as action, given that they

[2] Consider among others Attridge (2015), Spolsky (2015).

do not resemble our commonsense understanding of what an action is—namely, a physical movement. Second, Gaynesford raises the question about the 'deed done' via the action contained in poetic utterance: are these things done once and for all (as when Chaucer dedicates his poem to the Lord), or are they done anew each time someone reads a certain poem? Third set of problems concerns the issue of agency: who in fact is doing the deed, the poet, the poem, the 'lyric subject' or some other theoretical postulate? With respect to the latter two questions, Gaynesford ultimately concludes that their theoretical implications relate to interpretations of individual poems, and do not amount to reasons to dismiss his theory. To answer the first question, he invokes a distinction introduced by Austin himself, between ordinary physical actions and the special nature of the act of saying something. Claiming that the poetic utterance falls under the latter category, he ultimately sees the problem of classifying poetic utterances as instances of an action as a question that should be considered within philosophy of action, rather than as a question pertaining to debates on poetry. Independently of whether or not such an answer is sufficient to silent those who might object to his approach, it is not altogether a mistake to say that Gaynesford should tell us more about his own understanding of action, given the complexities involved in the notion itself, particularly when introduced into aesthetic debates (Davies 2011). This is particularly so given the emphasis he puts on the notion of responsibility, and on the question of 'whether what was stated has been performed', which he poses as a criterion for the poem's success (as opposed to the question of whether what is stated is true). While interviews, diaries and other evidential support he invokes to support his theory might work for some poets, they do not necessarily account for many others, particularly those who are long gone.

Gaynesford's tactic of undermining Austin's disregard for poetry as non-serious is simple: since Austin does not discriminate between different kinds of poetic expression and claims on various occasions that "poetry is 'a use of language' which is 'not serious'" all we need to prove him wrong is find one instance of poetry that can be regarded as serious. Of course, the idea of "serious" and "non-serious" uses of language is a complex one since it derives from very vague and ambiguous use of the terms (see 42-68), but what Gaynesford is devoted to is to find (at least) one instance of poetry that can be "responsible, committed, and thus 'serious'", that is, that can be used to back up the claim that we can "do things with poetry" in the real word. According to him, responsibility can be of three sorts: pragmatic, aesthetic and ethical (107-8). It is important to notice that according to Gaynesford there is no responsibility without intention: "Did the person who performed this action really mean to do what in fact they did? Did they realize what obligations would be laid on them by doing this? Did they accept, consent to, or undertake these obligations? For if the answer to any of these questions is ‘No’, then we may refuse to hold the person responsible for what was done, or at least qualify their responsibility, at each of the three levels: pragmatic, aesthetic, ethical.” (109) Trying to give an answer to questions of this sort in relation to poems whose authors are long gone could be puzzling. How do we reconcile the temporality of the poem’s author with the atemporality of the work of art? If the poet had a certain intention at the moment of the utterance, that is, at the moment he penned it in the form of a poem, he could be held accountable at that moment, but the analysis becomes metaphysically dubious once the referent of the “I” in a poem is gone—provided we can agree to identify it with the author as Gaynesford does when he claims that “in successful poetic utterances, poets perform acts of responsibility and commitment” (114). On the other hand, if we do not identify it with the poet, then all talk about real-life commitment and responsibility becomes vacuous.

Gaynesford acknowledges that “some would deny that responsibility and commitment are ever possible in the particular context that is poetry” and defends his position once again claiming that we need only one good example of “serious” poetry: “To undermine [Austin’s position], we need not argue that poetry is always, or indeed usually, responsible, committed, and thus ‘serious’. We need only produce examples of commitment-apt utterances in poetry where there is a genuine attempt to make that commitment, and where that commitment is indeed made.” (110) The Chaucer-type phrases are thus introduced as indicators of responsibility. Still, this is not an unequivocal answer since it leads to the question: who is the agent? Gaynesford acknowledges that there is no simple answer to this question and proposes a case by case approach—every poem will provide a new challenge: “(…) the claim that poetic uttering can count as a form of action, a speech action, raises difficulties. But none of these difficulties amount to objections to the overall claim. Rather, they set an agenda for the interpretation of specific poems, a list of questions that interpretations must resolve to count as satisfying. And this agenda proves an essential device. For where these difficulties arise, they direct the attention to the very issues that the poem itself is trying to raise.” (106)

Gaynesford’s strategy is to inspect every poem, or perhaps even every verse in a poem, to find a proof of commitment on the part of speaker to the content of the poetic utterance. If we find one example of a committed speaker in a poem we have falsified Austin’s claim that all poetic utterances are non-serious. If we concede this point, agree with Gaynesford’s interpretation of Austin’s view of poetry (see ch. 3) and find one or more poems that satisfy the criteria of his action-oriented approach to poetry, we can still wonder if the fact that we can analyze only a small portion of poems using this adapted speech-act framework does not point to a weaknesses of the proposed approach. One counterexample is enough to falsify a theory, but we need more than one example (or a few of them) to build our theory. This is true especially if the theory we try to falsify is in fact not about the particular phenomena we focus on in our attacks on it, and Austin's theory is not about poetry.

Gaynesford's account could be bolstered by a more substantial account of the way in which composing a poem can be understood as an instance of action, except in the sense in which the act of writing is itself an instance of action—which, of course, is not what Gaynesford's account suggests or aims to establish. While in the Chaucer example it is unproblematic to recognize the deed done—the dedication of a certain poem to someone—some other examples that Gaynesford uses might not work quite as easy. Consider his treatment of Douglas Dunn's poem 'Arrangements':

And here I am, closing the door behind me,
Turning the corner on a wet day in March.

As Gaynesford argues, "the line-break acts like a corner to be turned, thus enabling the utterance to do precisely what is says" (101). However, it seems strange, if not utterly impossible, that an act of saying does the job of turning the corner, independently of the line break. In other words, it is the act of walking that makes one turn the corner, not the act of saying that one is turning the corner (or an act of inserting the line-break in the appropriate place in the poem). The most that these two lines do is describe what the poet is doing, but they are not doing the deed (i.e. the act of turning the corner) itself. In that sense, even if 'what is stated is done', this still does not count as an instance in which poetic utterance has in fact committed any kind of action (other than that of describing). The question then remains for the reader to decide whether this is an instance of an ill-chosen example, or whether we should demand more in terms of criteria which turn *some* (as Gaynesford rightly emphasizes) poetic utterances into actions.

Perhaps such criteria would be available if more was said about questions two and three identified above. Namely, if poetic utterances are a form of action, what precisely is the deed done or brought about via these actions? In some cases, as with Chaucer, it is the one of dedicating a poem to someone. However, even assuming the plausibility of categorizing such poetic utterances as a form of action, what are the implications of that categorization for our understanding of poetry? In other words, does the fact that some poems are dedicated to someone, or that some poets invoke the help of the Muse or manage to perform some such action via their verses, justify the acceptance of the 'poetry as action' paradigm, or does it merely point to the (another) interesting way in which language works in *some* poems?

An answer to this question is, arguably, suggested by Gaynesford's interesting analysis of Shakespeare's sonnets. This analysis gives the impression that what is in fact done, the action that is triggered by the composition of a poem, is better located in the workings of the poem, i.e. in the way in which it initiates (philosophical) reflections in the reader. Bluntly put, what the poet does, on this interpretation, is not only dedicating his poem to someone, but a more substantial act of causing his readers to undergo certain experiences. After all, poetic encounters leave us with the sense of having undergone some kind of emotional and intellectual experience—for example, that of recognizing and appreciating, potentially even engaging with, the sceptical worries underlying Shakespeare' sonnets, or, to suggest our example, of sensing the pain and disappointment of a speaker who urges us 'Never to give all the heart' in Yeats' famous poem of that title, and then of considering whether one would indeed renounce the possibility of passionate love in light of the poem. This line of thinking about the attunement is in line with the criteria Gaynesford himself emphasizes: in order for poetry to be serious, poetic utterances have to exhibit commitment and responsibility. In other words, one has to be capable of doing what one says. More elaborately, "those responsible for poetic utterances must be able to count as such in a deeper sense than mere causal efficacy. It must be possible and actual for them to commit themselves in saying what they do. Hence it must be possible and actual for them to be, and to be held to be, responsible in what they say." (110)

This criterion will naturally raise the bar for what counts as *serious* poetry, i.e. which instances of poetry might count as serious (even if it does not help us account for what is for something to be poetry). When Yeats (if indeed Yeats it is, rather than the lyrical subject) urges us 'Never to give all the heart', and enlists rather persuasive arguments for such a statement in his poem, are we to take him seriously, or are we to enjoy the particular way in which the rhythm and rhyme work together to make this poem an aesthetic delight? Would he himself commit never to give up all the heart? Would he, let us wonder, repudiate his own advice had he but had a chance for happiness with his long desired Maude Gonne? Another problem that arises from embracing the 'responsibility and commitment' criterion relates to the fact that Gaynesford's account presupposes some type of intentionalism on the part of the poet to do certain acts—namely those for which he is willing to take responsibility. But, as numerous critics of intentionalism have pointed out, it is not necessarily so that poetry is to be considered, appreciated and evaluated according to the standards provided by the intentionalist framework.

Some poets of course do commit and can be held responsible for what they are saying. To consider their poems as an instance of an action, rather than as true (or false) array of statements referring to the real world, is a plausible move, if by action one has in mind a kind of intellectual activity that takes place in the readers' mind in the 'afterlife' of a poem (as Peter Kivy might put it (Kivy 2006)), or that inspires poets to turn to particular issues and write about them. When Kant talks about poetry 'animating the mind' (Kant 2000, for a discussion see Šustar and Vidmar 2016, Vidmar 2018, Vidmar forthcoming), he might think of some such understanding of the ways in which poetry does things to us, in addition to moving us via the sheer power of its aesthetic qualities. Consider much of religious poetry or various instances of metaphysical poetry. Robert Frost's repeated questioning into the moral status of natural creatures and men's relation to the world, satisfy, we think, not only the claim that poetry can be an action, but exemplify a poet committed to that what is stated in his writings and willing to take responsibility for such actions, even if not always using Chaucer-type of utterances. At best then we can conclude that, as usually the case with philosophical theories, Gaynesford's account works for some, but not for all poetry, and does not cover all instances of poetic creation.

What then to conclude regarding the connection between poetry and philosophy? Certainly, Gaynesford has a point in stating that the attunement approach challenges our understanding of both, poetry and philosophy. To understand the way in which Shakespeare manages to develop a view on the passing of time or to envision sceptical concerns makes a demand on scholars to reconsider ways in which philosophy can be conducted, as well as the limits of poetic engagements. On the other hand, philosophers such as John Gibson or Peter Lamarque might nevertheless insist on the futility of attunement, each for his own reasons. Gibson could argue that even if poetry is a form of action, its ties to philosophy are not established, given that the two disciplines do not entwine but remain separated by the mere diversity of their methods. Lamarque, himself a fervent opponent to approaching poetry from the standpoint of the truth debate, might argue that attending to the way in which Shakespeare develops a sceptical view is not to be evaluated by philosophical but literary/aesthetic criteria.[3] Consequently, nothing much is gained in terms of developing a more elaborate account of poetry *generally*, by appreciating philosophical considerations of *some* poems. It is not clear, to us at least, that Gaynesford's account would seem convincing to someone who shares Gibson or Lamarque's concerns. What is convincing though is his plea for taking poetry seriously and to continue analysing its ties to philosophy.

References

Attridge, D. 2015. *The Work of Literature*. Oxford: Oxford University Press.

Davies, D. 2011. *Philosophy of the Performing Arts*. Oxford: Willey.

De Gaynesford, M. 2017. *The Rift in the Lute: Attuning Poetry and Philosophy*. Oxford: Oxford University Press.

Gibson, J. 2017. "What Makes a Poem Philosophical?" In K. Zumhagen-Zekple and M. LeMahieu (eds). *Wittgenstein and Modernism*. Chicago: University of Chicago Press.

[3] In arguing this, we are taking cues from Gibson 2017 and Lamarque 2009.

Gibson, J. E. 2015. "Introduction." In his *The Philosophy of Poetry*. Oxford: Oxford University Press.

Kant, I. 2000. *Critique of the Power of Judgment*. Ed. by Paul Guyer, transl. by Paul Guyer and Eric Matthews. Cambridge: Cambridge University Press.

Kivv, P. 2006. *The Performance of Reading: An Essay in the Philosophy of Literature*. Oxford: Blackwell.

Lamarque, P. 2009. "Poetry and Abstract Thought." *Midwest Studies in Philosophy* 33 (1): 37-52.

Ribeiro, A. 2009. "Toward a Philosophy of Poetry." *Midwest Studies in Philosophy* 33 (1): 61-77.

Spolski, E. 2015. *Contracts of Fiction*. Oxford: Oxford University Press.

Šustar, P. and Vidmar, I. 2018. "Beyond Formalism in Kant's Fine Arts." In V. Waibel (ed.) *Freedom and Nature. Proceedings of the XII International Kant Congress*. Berlin: DeGruyter.

Vidmar, I. 2016. "Challenges of Philosophical Art." *Proceedings of the European Society for Aesthetics* 8: 545-569.

Vidmar, I. 2020. "Kant on Poetry and Cognition." *Journal of Aesthetic Education.*

Croatian Journal of Philosophy
Vol. XVIII, No. 54, 2018

Book Reviews

Jennifer Nado (ed.), *Advances in Experimental Philosophy and Philosophical Methodology*, London: Bloomsbury Academic, 2016, 179 pp.

Philosophical methodology has rarely been scrutinized and a subject to various opposing accounts as much as in the last decade. One of the reasons for this are challenges raised by the naturalistic movement of experimental philosophy (xphi), which offered a negative perspective and many critiques of, generally speaking, the dominant view in contemporary philosophy about what philosophy is all about, and the significance of its distinctive method, i.e. intuitional methodology. By using methods of empirical sciences and conducting numerous researches relevant to various disciplines in philosophy, experimental philosophy challenges the overly reliance on the method of cases and intuitions as a source of evidence. As time progressed, initial experimentalist's challenges required several modifications as they received a lot of criticisms by philosophers who endorse intuitional methodology as well as those who are skeptical of it. So this volume is about experimental philosophy in relation to intuitional methodology, and attempts of "reexamining its roots—to articulate just what the targets, aims, and methods of experimental philosophy really are" as Jeniffer Nado states in the introductory part (4). And each of the contributing articles gives, in one way or another, a new perspective on how experimental philosophy is to be understood, or in what direction it should advance. In this fashion they provide a useful insight into the metaphilosophical issue from experimental philosophy's point of view. This volume is one in the series *Advances in Experimental Philosophy* edited by James R. Beebe and in many levels brings insightful perspective on the currently highly debated topic in metaphilosophy, that of appealing to intuitions.

In the first article of this volume, Jonathan Weinberg, discusses the relation between the two important epistemological and methodological notions: *reliability* and *trustworthiness*. The latter is especially important in the light of Weinberg's new perspective of how experimentalist's challenge should precisely be formulated. Weinberg starts his discussion with questioning the hypothesis that it is reasonable to accept some source of evidence only on the basis that it is a reliable one. This would be true, Weinberg continues, if reliability is the "main determination of the methodological trustworthiness" (12). But since any degree of reliability less than perfect is consistent with the high degree of untrustworthiness, Weinberg argues that the notion of baseline reliability is methodologically inadequate. That means that even if intuitions are regarded as a reliable source of evidence, they are, methodologically speaking, untrustworthy. For one thing, the weakness in our inferential recourses can transform a highly reliable source to an inadequate one, and for another, intuitions do not have sufficient power to enable us to decide between two competing theories. To furthermore support his thesis, Weinberg discusses some theoretical implications of the current philosophical practice. One of them is the high vulnerability of philosophical theories to counterexamples where it is enough just to find one such example to overrule the theory, considering that this is not standard procedure in other sciences. For this reason Weinberg investigates the possibility of a different philosophical methodology. He proposes that we should ask ourselves whether philosophical truth must have exception-intolerant form and, consequently, whether we should put more weight on methodology that is exception-tolerant. Even if we decide that modally strong claims—such as "knowledge is…", where "is" is an identity claim—are worthy of philosophical pursuit, there are plenty theoretical results that are of value in achieving in philosophy in the exception-tolerant manner. The example of that are generic claims, as one such claim in epistemology, e.g. knowledge is justified true belief, is very useful peace of epistemological lore, according to Weinberg. And since the classical philosophical method of appealing to intuitions is not very useful in testing rival philosophical generics, Weinberg sees precisely this area as an appropriate place for experimental philosophy. It can give us tools for measuring the preference of one theory over the other, which are not just "hand-waving, it-seems-to-me kinds of ways" (29). According to this view, experimental philosophy could take the role of cleaning up philosophy's data set.

In "How to Do Better: Toward Normalizing Experimentation in Epistemology", John Turri is reviewing five cases where philosophers—or to be precise epistemologists—have deeply mischaracterized the "commonsense epistemology", conception they very frequently appeal to. For instance, epistemologists, almost unanimously advocate the idea that knowledge requires reliability and that this is a matter of common sense. But when this hypothesis is put to test, results show that knowledge judgments are insensitive to the information about reliability. Turri conducted a survey where participants, typically in similar percentage, attributed knowledge to both reliably and unreliably gained processes. He found the same results in the cases of contextualism, epistemic closure principle, truth-insensitive theories of justification, and knowledge attribution in "fake barn" cases. In each of these cases epistemologists typically argue that their proposed theory is "intuitive", "has basis in ordinary language", or that it is "a defining feature of commonsense epistemology" (40). But when tested, subjects typically do not respond as theory predicts. Turri concludes that the standard practice in analytic philosophy is to rely on "introspection and anecdotal social observation to characterize commonsense epistemology" (45), and that this has two potentially significant implications. First is a negative perception of the contemporary academic philosophy where people are suspicious of the possibility that important philosophical questions can be answered from the armchair. Second relies on the fact that people cannot relate to judgments that philosophers treat as obvious, intuitive or commonsensical (e.g. judgment that the "brain-in-the-vat" and normal human are equally justified in their beliefs, or that the agent in the "fake barn" case does not have knowledge). The role of preliminary experiments conducted by experimental philosophy can help to avoid these mistakes and thereby put researchers on more promising paths by avoiding the false start.

The move from talking about the main experimentalist's target, i.e. philosophical intuitions as part of what makes philosophy methodologically unique, to the talk about thought experiments, where intuitions are generated is proposed by Joshua Alexander in his article "Thought Experiments, Mental Modeling, and Experimental Philosophy". He thinks that this should be done by considering two dominant approaches to thought experiments in the philosophy of science: the "argument view" and the "mental model view". The underlying idea behind the first view is that thought experiments are nothing more than colorful arguments and that they can be reconstructed as premises and assumptions leading to the conclusion. According to the second view, thought experiments are not solely arguments because narrative of the thought experiments allows us to "mobilize cognitive recourses that would not otherwise be available" (58), in terms of manipulation of mental models in our imaginations. In other words, without the narratives in thought experiments, our ability to arrive at the conclusion they are intended to support would be compromised. According to Alexander, placing a philosophical cognition in the center of the debate along the suggested lines actually makes experimental philosophy, as an empirical study about philosophical issues, more important rather than less. To clarify this thesis, Alexander discusses one of the most controversial claims in experimental philosophy, namely the claim that people think differently about the narratives used in thought experiments. This is what he calls the "narrative incompleteness" problem, according to which many details in fictional narratives are often left out for the reason to be as less distractive as possible. Now, even though some opponents of experimental philosophy would argue that this feature of fictional narratives shows that there is no philosophical disagreement but instead that people simply have different fictional narratives in mind, Alexander claims that this should not be understood as a critique of experimental philosophy. It rather underscores the relevance of experimental philosophy because it investigates how people think about fictional narratives used in philosophical thought experiments, that is, to what information used in narratives people are responding. To conclude, by reframing the discussion in terms of thought experiments instead of intuitions, Alexander is maintaining that arguments against experimental philosophy could be reinterpreted in a way to actually support a need of experimental philosophy.

The only paper in this book that does not examine prospects of experimental philosophy in a positive way is "Gettier's Method" by Max Deutsch. He aims to revisit the broadly endorsed metaphilosophical view—also endorsed by the experimentalists—that analytic philosophy employs method of cases and that intuitions are essential part of this method. As can be extracted from Deutsch's paper, there are two interpretations of the "intuition-view" that are under his attack:

(i) Gettier cases are examples of appealing to intuitions as evidence, and
(ii) Gettier cases are examples of both arguments and appealing to intuitions as evidence.

Concerning the first interpretation, Deutsch argues that Gettier does not appeal to intuitions, and that since intuitions play no role in his argument against the traditional JTB theory of knowledge, his thought experiments are not examples of the method of cases. Deutsch's reasoning is as follows. (I) Gettier nowhere uses the term "intuition", and nowhere argues that we should accept his cases on the basis of intuitiveness. And the possibility that Gettier might appeal to intuitions implicitly is rather weak, according to Deutsch. Furthermore, (II) Gettier is not vague about the justification for his conclusions, and provides an explicit argument stated in his first case as follows: even though Smith believes truly and with justification that the man who will get the job has 10 coins in his pocket, his belief is merely lucky one and does not amount to knowledge. Regarding the second interpretation that Gettier cases are examples of both arguments and appealing to intuitions as evidence, Deutsch does not undermine its possibility, but insists that there is no evidence to suggest that such a possibility is actual. Even more puzzling for Deutsch is what he calls the "usual view", according to which Gettier does presents argument against the JTB theory, but he does not present argument for premises of this argument. Instead, as this view suggests, these premises are supported by intuition alone. Deutsch argues that explicit argument for the conclusion that Smith does not know that the man who will get the job has 10 coins in his pocket, namely, the presence of luck, qualifies as a good reason for denying Smith's knowledge. And concludes that it is a mistake to understand Gettier cases in a way that he intended for intuitions to reveal the falsity of JTB theory of knowledge. The further reason Deutsch discusses of why we should reject the view that thought experiments are about appealing to intuitions is that the post Gettier literature proceeded in an entirely intuition-free way (e.g. Michael Clark (1963), Alvin Goldman (1967)).

The problem for Deutsch's view could potentially be an interpretation that the order of explanation goes the other way around, namely, via abduction Gettier is arguing that the anti-luck premise is the best explanation of the truth of the conclusion that Smith does not know. This could pose a problem for Deutsch's position only if the anti-luck condition is intended to be fully abductive, and he thinks that this is extremely unlikely. One reason is that at the time of publishing Gettier's article, it would be highly controversial and unorthodox to take conclusions as granted in order to abductively argue for the anti-luck condition. It is more likely, according to Deutsch, that Gettier intended it the other way around. Additionally, the so-called producer-consumer distinction serves as a further reason not to accept Gettier cases as paradigm examples of the method of cases. As Deutsch sees it, Gettier himself could not use intuitions as evidence since the process of constructing thought experiment is anything but "passive sort of cognizing characteristic of intuiting that something is so" (85). And even thought, we as consumers, might experience intuitions about his examples, this is irrelevant for its evidential status since Gettier construed his cases as counterexamples, and presumably had evidence for it before we get the chance to read them.

In the next article titled “Intuitive Diversity and Disagreement” Ron Mallon considers a subset of critiques against the experimental philosophy, specifically, the subset that argues that, even though Platonic armchair method (i.e. the method consulting a priori intuitions about general philosophical truths) is a bad methodology, it is not wildly employed by philosophers. And this subset of critiques offers alternative explanations of what exactly philosophers employ in such cases. Mallon’s aim is to argue that experimental philosophical challenge—or at least one version of the challenge, namely the argument from disagreement—poses the same problem for these alternative interpretations of the philosophical method as it does for the Platonic armchair method. This is because Mallone holds the following two assumption: (i) the challenge “need not depend on attacking a distinctively Platonic armchair, or on any eccentric psychological construal of the relevant mental states” (108), and (ii) intuitions “pick out the sorts of seemings or judgments involved in our target cases” and also “behavioral manifestations of those judgments produced in response to philosophical thought experimental surveys” (100).

One of the alternative interpretations under Mallone’s critique is the suggestion that philosophers do not actually appeal to intuitions as evidence. One example of this alternative interpretation is presented in the previous section when discussing Deutsch’s view. First of all, Mallone rejects the underlying idea of this alternative interpretation that just because an author gives an argument for the proposed conclusion, it follows that author’s spontaneous judgment that *p* plays no evidential role. According to Mallone, this is not a valid inference, for both spontaneous intuition as a source of evidence and reasons why intuition is considered to be true can be held at the same time. And this is, in Mallone’s view, supported by many thought experiments where it is obvious that they are not to be understood as pure arguments, because in order to be valid, they must be supplemented with substantial assumptions about topics under investigations (e.g. assumption about the nature of knowledge). But even if we allow that philosophers do appeal to intuitions in their arguments, critics would further argue that they need not to do so, and thus variability in intuitional judgments would no longer pose a problem for philosophy. And at this point, Mallone shows in what way this alternative explanation does not avoid the problems of the argument from disagreement. Namely, experimental philosophy criticizes “*actual* rather than *possible* practice” (115), and thusly still presents the problem for philosophical practice. The other alternative interpretation that Mallone considers in his paper is the mentalist approach, which takes that intuitions do not reveal some abstract reality as the Platonic armchair approach does, but rather facts about human concepts, or some other psychological mechanisms that produce intuitions. But nonetheless, mentalist approach is also affected by the argument from disagreement, because its proponents are interested in shared concepts. And whether some particular concept is common cannot be revealed from the armchair.

Jenniffer Nado further develops the idea of “reexamining roots” in the paper “Intuitions and the Theory of Reference” she coauthored with Michael Johnson. The general idea that they develop is that experimental philosophy is especially relevant in the theory of reference, but reasons for its relevance cannot be extended to other fields of philosophy. To show this, they focus on the particular experimental study conducted by Edouard Machery et al. where they find that cross-cultural differences in responses to Kripke's Gödel case undermine the viability of the intuitional methodology. Nado and Johnson argue that reports in such surveys is primary methodology in the theory of reference, but do not show, as Machery et al. claim, that relying on intuitions is a bad methodology. The reason is that such reports (i.e. judgments about cases) are "instances of speakers applying terms to things that have been generated under controlled conditions to test the predictions of different theories" (148). And cases of people applying terms for things are primary data for theory of reference. They furthermore argue that this reason is not straightforwardly applicable to other fields, since the correct application of terms, such as "time" and "consciousness", depend on some extra-linguistic facts that are not easily accessible, and therefore intuitions about those terms would be of little evidential use. So, the assessment of intuitions as a source of evidence will vary from field to field and consequently, so will the relevance of experimental philosophy.

In the last paper of this volume, "Intuitive Evidence and Experimental Philosophy", Jonathan Ichikawa claims that experimental studies are relevant for philosophical methodology, but only in the limited sense. His account of intuitions, that is intuitional methodology, is that it is a mischaracterization of philosophical practice to claim that intuitions are used in a central evidential way. But he also argues that this fact alone does not make experimental studies redundant (which is the usual stance for someone who denies evidential role of intuitions). Namely, he agrees with proponents of experimental philosophy that their surveys and interpretations of those surveys do not, in any clear way, depend on the assumption that intuitions have an evidential role. This is defended from the standpoint that empirical investigation of intuitions can be relevant for philosophical methodology even though they do not play evidential roles, since evidential role is not the only role of epistemic significance. However, Ichikawa thinks that this alone is not enough to defend experimental philosophy. He argues that even though experimental philosophy survey's results do not essentially make use of intuitions the same does not hold for their analysis. In other words, the replacement of the term "intuition" with any other non-problematic term in their analysis cannot be done straightforwardly. And this is where Ichikawa sees the biggest challenge for the defense of experimental philosophy, although not as big to make it irrelevant. For example, proponents of experimental philosophy often claim that intuitions are susceptible to order-effect which makes them not suitable as evidence. Even under the assumption that intuitions are not to have an evidential role, the fact that they are so susceptible should be the reason to doubt one's ability to rationally respond to the available evidence, and seek guidance how to proceed thereafter. To sum up, philosophical biases are epistemically relevant, and it is worthwhile to engage in attempt to detect them.

ANA BUTKOVIĆ
University of Zagreb, Zagreb, Croatia

Miranda Fricker and Michael Brady (eds.), *The Epistemic Life of Groups,* Oxford: Oxford University Press, 2016, 272 pp.

Can groups hold, revise and reject beliefs? Are collective doxastic attitudes reducible to what is believed by individual members or do they presuppose some additional joint commitment? How can we resist the sway of social stereotypes when assessing others as moral and intellectual agents? Are emotions always a hindrance to epistemic goals and is the apparent conservatism of scientific groups indeed a deviation from usual collective behavior? Departing from analytic epistemology's traditional focus on individual agents who operate in something akin to a social vacuum, this volume explores the epistemic features of group agency. Its contributors inquire, for instance, to what extent collective processes of attaining and revising beliefs can be equated with their individual counterparts, and whether belonging to a particular intellectual environment can generate morally corrosive prejudice. The volume consists of four thematic clusters concerning epistemology as such, moral epistemology (understood as the practice of attaining beliefs about actions related to morally valuable outcomes), politics and science. However, portraying the work as a handbook one should recommend to a novice would—despite its stated introductory aim—be somewhat misleading, as it presupposes considerable familiarity with prior discussions on testimony, epistemic injustice, deliberative democracy, assertion, Kuhnian philosophy of science, and like. The actual importance of certain articles, such as Miranda Fricker's apt revision of the overly optimistic approach to implicit biases she had argued for in her earlier works, can only be fully appreciated if one is well-acquainted with recent trends in social epistemology. Taken as a whole, nonetheless, *The Epistemic Life of Groups* presents the reader with a range of engaging topics that merit further attention. For the love of simplicity, I will remain true to the volume's structure in offering brief comments on each essay.

Epistemology. Sandorf Goldberg opens the first section with the claim that criteria for considering an assertion proper depend on the intellectual community under whose auspices it is uttered. Within the context of some conversational group riddled with such pervasive disagreement that hopes for attaining knowledge dwindle, assertions are proper as long as they serve the group's informational purposes and can be reasonably expected to be understood by other members. Although Goldberg intends to preserve the propriety of philosophical discourse despite the community's continuous dissent on central issues, his immediate acceptance of contextualism seems the overlook the stronger case that philosophical assertions often hinge on objective standards—such as logical validity in narrowly theoretical domains and congruence with experimental findings when discussions veer closer to cognitive and social science—which render certain statements more pertinent to knowledge than others. Instead of exonerating philosophical discourse, this decision to trade the more demanding epistemic norms of knowledge or empirical adequacy for reasonable in-group intelligibility forces us to concede that collectives which are usually considered epistemically irresponsible—such as science deniers or conspiracy theorists—actually do satisfy a more forbearing epistemic norm, given that their assertions are entirely in sync with other members.

Miranda Fricker proceeds with a sensible review of her work on epistemic injustice and recognizes that individuals are seldom able to fully overcome the biases they had internalized by growing up in a prejudiced society. To what extent, then, to these cognitive constrains pardon us from blame for wrongful epistemic conduct? Although implicit biases—as they run counter to our consciously held values and therefore cannot be considered intentional—aren't conventionally culpable, this falls short of excusing our behavior. Making use of Bernard Williams' definition of agent-regret as the appropriate response of someone who had experienced a case of moral bad luck, Fricker argues that otherwise conscientious perpetrators of epistemic injustice should regret their misconduct, reflect upon their prejudiced beliefs and encourage institutional measures which will prevent their peers from repeating similar mistakes.

After this brief foray into practical concerns, Hans Schmid wonders whether group self-knowledge is as groundless (meaning, as automatic and as non-inferential) as its individual counterpart. Albeit he first shows that Anscombe's criteria for individual intentionality—namely, first-person identity, perspective, commitment, and authority—aren't intuitively compatible with the collective model, Schmid ends up concluding that genuine group belonging does require a strong sense of joint commitment and identification which render the idea of groundless group self-knowledge sensible (72).

Ethics. In the volume's first essay on moral epistemology, Elizabeth Anderson offers a rich account of how social moral learning—the collective acquisition of true beliefs about our ethical duties to others—may be obstructed if we indulge in sanitized interpretations of historical injustice. When entire communities agree on self-laudatory narratives of their previous moral excellence—which Anderson illustrates with the fact that slavery wasn't abolished due to the autodidactic moral learning of Western intellectuals, but, instead, because subalterns continuously sought human rights—they fail to acknowledge that only the disadvantaged have substantial epistemic access to the urgency of their problems (78). The historical facts that whites first envisioned a gradual abolition of slavery that would take decades and then offered freed blacks unlivable wages for working the same fields they had previously occupied as slaves make the problem quite salient. What Anderson appeals for is a kind of epistemic democracy wherein moral progress requires those in privileged social positions to recognize the humanity of their interlocutors and to, when crafting policy, take their experiences into account on terms of equality.

Michael Brady follows up with the original claim that emotions—both individual and group—can have epistemic value inasmuch as they direct our attention towards certain events and compel us to appraise whether they had warranted such an emotional response, thus promoting understanding. Arguing that individual emotions amount to group counterparts through emotional contagion and affective conformity, Brady concludes that shared dismay with social events makes groups inquire about what is indeed going on and, consequently, may encourage greater transparency from governing bodies (109). He does note, however, that misinformed group emotions can cause severe epistemic harm and hence require situational assessments.

Next up, Glen Pettigrove wonders whether the propositional model of revising beliefs within groups can explain how moral communities change complex opinions with holistic content (121). Heavily drawing on Margaret Gilbert's collective epistemology, he uses the example of the Presbyterian Church to show that revisions of moral knowledge do not arise when members merely replace one proposition with another, but instead require shifts in comprehensive—or holistic—moral doctrines.

Politics. Fabienne Peter inquires whether democracy can be justified in the light of its epistemic value alone, rather than by appealing to practical concerns. Simply put, if we can resolve political matters by making a correct decision, presupposing that there is an objectively correct choice to be made, then democracy is legitimate inasmuch as its decision-making procedures reliably produce such outcomes. The problem here lies in what she calls "the authority dilemma" (134). As long as there is a relevant third-person authority—say, an expert in some field—who is particularly knowledgeable about a matter of collective interest, aggregating the opinions of comparative laypeople will not seem like an advisable route to social policy. Peter first presents a case for deliberative democracy, which stresses the importance of exchanging reasons and acknowledging plural perspectives by way of public debate, in place of mere aggregation or majority voting. Next up, assuming that certain questions—such as highly contested, theoretical and ideologically laden issues—do not entail a procedure-independent truth, she concludes that the deliberative process is in itself epistemically valuable because it sensitizes agents to different opinions. This line of reasoning leads to the obvious conclusion that we should only entrust decision-making to democratic collectives in matters lacking an objective third-person authority (149). What remains to be explored is the precise domain of such purely subjective topics. Peter's portrayal of minimum wage policies as a subject that—although it is undoubtedly a matter of public dissent—requires no expert knowledge could be contested, so future discussions might benefit from a more careful distinction.

Stephanie Collins and Holly Lawford-Smith proceed by inquiring about the transfer of duties between individuals and states. This process, in their view, includes several epistemic components: individual members recognize their country in compelling it to discharge duties on their behalf, the state acknowledges its individual members by distributing smaller duties (such as taxes), members intentionally participate by fulfilling their obligations and both parties engage in bidirectional transfers of knowledge concerning their ethical demands (160). Individuals are, moreover, only justified in transferring their duties to the state if they can reasonably believe that it will truly act on their behalf.

In the final essay on politics, Kay Spiekermann turns to behavioral economics in explaining how agents tend to ignore—or distort—readily available evidence about the ethical opacity of their actions. Having identified four types of "moral wriggle room" (182) wherein agents deliberately avoid facts which entail moral obligations, convince themselves that moral norms are more lenient than might seem or deceive others about the scope of their rights, Spiekermann locates them in discriminatory practices of "white ignorance." In this sense, whites tend to embrace faulty beliefs (such as the stances that freed black slaves had equal opportunities to whites or that black communities are only marginally disadvantaged) which diminish normative constraints on their behavior. Spiekermann does imply, however, that encouraging agents to voice their ethical values—and thus identify with them—can eliminate self-serving biases by rendering cognitive dissonance more explicit. Echoing Fricker's work on internalized prejudice, he concludes the essay by admitting that "it remains unclear whether training individuals to resist self-serving biases can succeed" (188).

Philosophy of science. James Owen Weatherall and Margaret Gilbert introduce the final section by combining Gilbert's seminal work on joint commitment and Thomas Kuhn's description of "normal science" (203). Arguing that group membership imposes certain responsibilities on its members, including a heightened sense of identification with the collective and a disregard for outliers' opinions, they use this joint account to show that the string theory community's "unusually" dogmatic behavior in contemporary physics only serves to confirm Gilbert's model. The upshot here is that the apparent epistemic irresponsibility of scientific communities—assuming that propensities to dismiss all opposing evidence and believe desirable results without checking don't live up to most methodological standards—isn't an occasional deviation from proper conduct, but a natural feature of joint commitment. This conclusion may serve as a sound basis for exploring common constraints on scientific progress.

Torsten Wilholt closes the volume by attempting to locate the source of trustworthiness in collaborative scientific research. The problem here is how one can assess whether a scientific collective is worthy of trust without appealing to traditional indicators of reliability such as institutional reputation. Noting that the social organization of scientific work has become so diffuse that it is almost impossible to attribute trust by employing previous track records, Wilholt argues that researchers can rely on shared methodological standards (229). The choice of a particular methodology, moreover, usually entails a trade-off between the reliability out its results—both positive and negative—and its power, or the number of generated results. Given that the dilemma between a small number of accurate results and more fecund, but less reliable research cannot be resolved by appealing to truth, researchers ought to attune their choices to the gravity of the matter at hand (233).

Regardless of the breadth of covered topics and the considerable quality of individual essays, the volume is better described as a compilation of different approaches to both collective epistemic agents and their individual members, than as a comprehensive introduction into collective epistemology. Having said this, an informed reader will surely find Fricker's and Brady's editorial work deserving of close philosophical scrutiny, and we can hope that this new territory will generate fruitful developments in the domain of collective and social epistemology.

HANA SAMARŽIJA
University of Zagreb, Zagreb, Croatia

Stuart Glennan, *The New Mechanical Philosophy*, New York: Oxford University Press, 2017, xi+266 pp.

The development of what is now known as the New Mechanical Philosophy started in 1990s, achieved groundbreaking status during the beginning of the millennium, and established itself as one of the most discussed topics in contemporary philosophy of science over the last decade. As Stuart Glennan, the author of the book *The New Mechanical Philosophy* points out, the name does not designate a school of thought or a movement, but rather a group of philosophers who revived the philosophical talk of mechanisms and their importance across all scientific fields. Indeed, various new mechanist philosophers share different views about mechanisms and their nature, with Glennan offering one such personal account of "how things hang together", to use his own phrase, in the form of a summary on the work done in the field. The book has eight chapters out of which six are dedicated to the ontological problem of what mechanisms are, with the other two chapters discussing new mechanism in general, and the problem of explanation. With the language accessible to philosophers and scientists alike, *The New Mechanical Philosophy* provides an excellent overview of this novel approach to thinking scientifically, both as an introduction to the topic, and as a systematic reference for those well informed in the field.

In Chapter one, titled "What Is the New Mechanical Philosophy", Glennan explains the motivation that drove the need for a new mechanical approach, its roots, and its peculiarities. New mechanists distance themselves from the traditional approach of "craving for generality" which Glennan sees as a perceptual and methodological hindrance that has plagued scientists, philosophers, and common folk alike. Although the roots of this philosophical approach can be found in as far as Democritus' atomism, seventeenth century mechanism, and again in 1960s, there is much 'new' in New Mechanism. Most notably, instead of talking about laws and generalizations, new mechanists have shifted their research to talking about mechanisms, and instead of talking about theories they have shifted over to talking about models. Glennan does not, however, elaborate this rejection of generality at length but simply designates it as an approach that is too far from the reality of the world; a reality that is, in new mechanistic view, first and foremost particular. For this reason, the main approach in the following chapters is an ontological one insomuch as it focuses on defining what these particulars all around us, i.e. mechanisms are how we can properly represent them via models.

Chapter two, titled "Mechanisms", explores the ontological status of mechanisms by discussing its constituents. Following the main thesis of the book, Glennan defines characteristics of a "minimal mechanism" in order to show that the talk of mechanisms is a common denominator of all scientific fields, explaining that "A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon"—a definition flexible enough to be applied to most of scientific explanations. One important aspect the author often discusses is stability; more specifically, stability of entities' properties and boundaries, and stability of mechanical production. These stabilities enable the scientists to use mechanical approach in order to explain regularities, however, as Glennan warns, a defined mechanism can *never* be taken as a strict law—its reality is always, and should be taken, as a particular. Each of the elements of mechanisms have been discussed over the last few decades, so this chapter functions as a concise introduction to prepare the reader for the discussions in the book that are yet to follow.

In Chapter three, titled "Models, Mechanisms, and How Explanations", the author elaborates on how we can represent particular mechanisms via models, which provide a type of general explanation, and which can represent more than one phenomenon. Models are still particular, and they are not to be confused with theory which is, in its abstractness, only a "toolkit for building models", or with laws, which are useful, but descriptions too idealized to be an accurate representation of particular mechanisms. In order to explain a phenomenon, Glennan argues, by explaining *how* it works (its underlying mechanisms), we will explain *what* it is. In order to prove the superiority of mechanistic explanations, this chapter introduces the reader with models as a midway between mechanisms as completely particular explanations, on the one hand, and theories and laws as completely abstract, on the other hand. This feature of models, as a certain level of generality, enables scientists to use them in order to explain various particular phenomena in a detailed and precise manner.

In the fourth chapter, titled "Mechanisms, Models, and Kinds", the author discusses abstract representations of particular mechanisms, and related problems. The aforementioned use of models proves useful even here. If, in our tendency to seek generalizations, we want to define kinds of mechanisms, the new would advise us to seek similarities between particular mechanisms in as detailed and broad way as possible, and then to construct a model as an abstract explanation that encompasses these particular instances. This process, Glennan warns, is not completely arbitrary. Although the scientist is bound by natural constraints, the type of the kind and the model to be constructed depends on their goals, resources, and interests. This "model first" approach, dubbed by Glennan, acts as a more down-to-earth approach which distances itself, just like new mechanism in general, from abstractness of traditional laws.

In the fifth chapter, titled "Types of Mechanisms", Glennan expands upon his initial definition of minimal mechanism in order to show the complexity and richness of types of mechanisms which, as he optimistically concludes, would offer a basis for interconnecting scientific talk of the phenomena. The author thus discusses elements relevant to classifying mechanisms, such as: types of phenomena, types of mechanical organization, types of etiology (how the mechanism originated), and stochastic nature of some mechanisms. One interesting idea expressed in the chapter, albeit not discussed in length, is Glennan's treatment of social sciences and phenomena. Following new mechanical approach, Glennan insists that abstract social concepts, like 'democracy' or "doctrine", are not entities on their own, but that they can only produce change and effect if explained by their constituting entities, such as individuals and their particular interactions.

In Chapter six, titled "Mechanisms and Causation", the author covers a wide variety of approaches to discussing causation as the origin of mechanical production. Indeed, the problem of causation has a long history, and in this particular chapter Stuart Glennan attempts at situating the new mechanical talk of causes within general philosophical framework. For example, he discusses the ways of explaining causes of processes, the problem of production and relevance, which will be elaborated in the subsequent chapter, the use of truth-makers, manipulation, and generalization. The author tackles these problems by invoking various philosophical conjectures, which performs a good task at providing the aforementioned philosophical context. One aspect that strikes the eye and makes a good case for Glennan's argument that new mechanical approach to causation offers some unique benefits, in his explanation that, in order to explain causation, we need not hold onto laws, but instead it is sufficient for a cause to only once produce a certain phenomenon in order for us to call it a mechanism, and explain it via models.

In the following Chapter, titled "Production and Relevance", the author provides a more personal account of the problem by arguing with various philosophers and expressing his own views. Some of the problems covered are Wesley Salmon's etiological explanation, types of mechanical production, the problem of irrelevant production, the problem of non-productive causation, the problem of causal (i)relevance, and the problem of the fundamental level of mechanisms. The latter problem, of the fundamental level of mechanical production is quite peculiar. One could rightly ask what is the right level of examining mechanisms if we are to be thorough and properly scientific? If we take it to be the atomic and subatomic level, as so called microphysicalism would advocate, we enter a domain of non-classical and indeterministic relationships. Glennan gives a nonconclusive answer to this problem, but one has to keep in mind that new mechanism allows for a certain arbitrariness in choosing the scope and relevance of the mechanisms to be examined, as is already noted in the fourth chapter.

In the last chapter titled "Explanation: Mechanistic and Otherwise" Glennan reiterates his position that mechanistic explanation of phenomena is but one of many scientific explanations, which continues his pluralistic line of thought from the introduction that the aim of the book is to show that mechanistic explanation is "useful" and worthy of further implementation and elaboration. In this particular chapter he discusses scientific explanation in general, contrasting the mechanistic explanation with "bare causal" and "non-causal" explanation, as the only one concerned with the question *how* we arrive from causes to phenomena production. It is interesting that in this chapter, and, indeed, the whole book, the author deliberately circumvents the question of truth, and instead talks about the utility and applicability of mechanisms and models.

In the "Postscript", which acts as a short conclusion of the book, Glennan expresses his hope that the book's ontological outline of mechanisms would inspire scientists to think more about "how things hang together" and to look at phenomena in a new way with new methodological tools that he laid out in this book. There is no doubt that the readers of this book will start noticing all the wonderful mechanisms around them in a new manner, as soon as they flip the last page.

MISLAV UZUNIĆ
University of Rijeka, Rijeka, Croatia

Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence*, New York: Random House–Knopf, 2017, 543 pp (eBook).

Max Tegmark is a Swedish-American physicist and an MIT professor who loves thinking about life's big questions, from cosmology to artificial intelligence (AI), and has wrote the book *Life 3.0: Being Human in the Age of Artificial Intelligence* published by *Random House* on August 29, 2017., in which he explores the future of life, technology and AI, and its relationship to human beings.

If you are wondering about the title of the book, let me take a minute to explain it to you. Tegmark divides life into three stages: *Life 1.0, Life 2.0* and *Life 3.0. Life 1.0* represents the simple biological evolution, where both the hardware and software are evolved rather than designed. The hardware and software are determined by its DNA and can't be changed in a single organism's lifetime but can gradually evolve over many generations. *Life 1.0* can only survive and replicate. Examples of *Life 1.0* are bacteria and other single-celled organisms. Humans, on the other hand, are examples of *Life 2.0,* life whose hardware is evolved, but whose software is largely designed. By hardware Tegmark means the body, and by software, our concepts, ideas and extended abilities such as language; all the algorithms and knowledge from our senses, thoughts and emotions that we use to process the information and then decide what to do — "everything from the ability to recognize your friends when you see them to your ability to walk, read, write, calculate, sing and tell jokes" (2017: 42). *Life 2.0* represents the cultural revolution, except it can survive and replicate, it can also redesign much of its software by studying, thinking, writing, joking or inventing new technologies. Finally, *Life 3.0* is the AI which can design and upgrade both its software and hardware. It can replicate itself from scratch and build new bodies relatively quickly from raw materials, plus, it can also learn from its surroundings, gather and store information, learn from the past to avoid mistakes, enabling it to advance enormously and to transform itself more directly and quickly than our creativity enables us to do. In this sense, *Life 3.0* is the master of its own destiny and, when created, will represent the technological evolution of life.

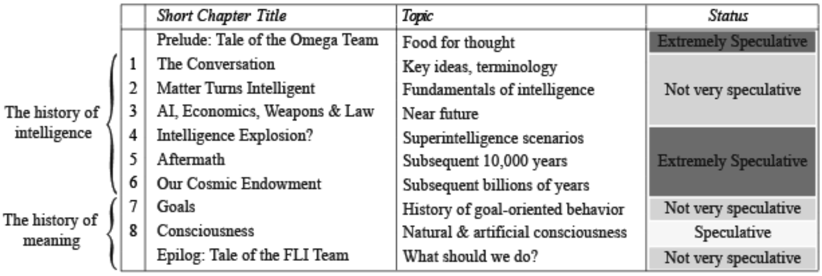
Tegmark says that the boundaries between the three stages of life are not so rigid. For example, if bacteria are *Life 1.0* and humans are *Life 2.0*, then a mouse could be between 1.0 and 2.0, let's classify it as *Life 1.1.* Although we might think that its software is also designed because of its ability of learning, the fact is that it can't develop a proper language to communicate or some other methods that can efficiently help him transfer the gathered knowledge to next generations. What a mouse learns during its life largely gets lost when he dies and the newborn has to learn from scratch by watching the elders. Similarly, we might argue that humans living in the 21st century are between 2.0 and 3.0. In fact, considering the current progress with prosthetics, artificial limbs and medicine in general, we are more like *Life 2.1* already. We can perform some hardware changes like replacing a missing body part, installing a pacemaker or a hearing aid, strengthening our *immune* system and curing many diseases with medications and so on, but we cannot design our body to be immensely different like having 4 arms, being 10 meters tall or having a thousand times bigger brain (45–46).

In summary he divides the development of life into three stages, distinguished by life's ability to design itself:

- *Life 1.0* (biological stage): evolves its hardware and software
- *Life 2.0* (cultural stage): evolves its hardware, designs much of its software
- *Life 3.0* (technological stage): designs its hardware and software.

Even if *Life 3.0* doesn't yet exist on Earth, it can arrive during this century and perhaps even during our lifetime. What will then happen and what this progress in AI will mean for us humans is the topic of this book, says Tegmark (46).

Now about the book. *Life 3.0: Being Human in the Age of Artificial Intelligence* is a relatively long book (almost 400 pages in printed edition), organized into eight chapters that take the reader on a journey that starts at the beginning of time, and describes the evolution of intelligence and technology in relation with humans. Here you can see the structure of the book given by the author (75–76):



		Short Chapter Title	*Topic*	*Status*
		Prelude: Tale of the Omega Team	Food for thought	Extremely Speculative
The history of intelligence	1	The Conversation	Key ideas, terminology	Not very speculative
	2	Matter Turns Intelligent	Fundamentals of intelligence	
	3	AI, Economics, Weapons & Law	Near future	
	4	Intelligence Explosion?	Superintelligence scenarios	Extremely Speculative
	5	Aftermath	Subsequent 10,000 years	
	6	Our Cosmic Endowment	Subsequent billions of years	
The history of meaning	7	Goals	History of goal-oriented behavior	Not very speculative
	8	Consciousness	Natural & artificial consciousness	Speculative
		Epilog: Tale of the FLI Team	What should we do?	Not very speculative

As the author points out, it's possible to skip some chapters because they are mostly self contained. For example, if you are not new to AI, skipping chapter 2 will get you right to the question "What does it mean to be human in the age of AI?" On the other hand, readers new to the AI field will get a nice introduction and terminology of the field in chapters 1 and 2 (74).

Tegmark begins his book with a fictional "what-if" premise, a very plausible but imaginative tale of the Omega Team, a corporate team of brilliant researchers who, with a commitment to helping humanity, secretly build an AI called *Prometheus*. With strong security measures, this superintelligent machine not only makes billions for its creators, but takes over the world and transforms it positively. Using this AI technology, the Omega Team accomplishes the most dramatic transition in history; eliminating all previous national power structures they create a world alliance and consolidate a single global power which runs the planet, ending state conflict; improving the quality of life, education and health; increasing the entire planet's standard of living and enabling life to flourish into the far future throughout the cosmos. This prelude can be read as a SF thriller but in fact it's much better than what Hollywood has come up with so far on this subject, and I believe it works very well as an introduction to the book because it's presented more like a non-fictional description of a business and political development rather than a sci-fi scenario. As the book progresses, the author occasionally includes fictional scenarios that fit the description of the Omega Team and Prometheus. This is genuinely thought-provoking and brings the technology and its human implications vividly to life. A great way to start the book and attract the attention of the reader (10–34).

After introducing the "Most Important Conversation of Our Time", giving the terminology and clearing some common misconceptions, in the second chapter of the book Tegmark gives a detailed description of intelligence, memory, computation and learning. He then discusses these qualities in the context of whether they are limited to humans (*Life 2.0*) or applicable to machines (*Life 3.0*) as well. The author gives an overview of intelligence from its origin, billions of years ago and through its development to the present days. He then explores the current state of research into machine learning and some breakthroughs in the field of AI.

In the third chapter, Tegmark goes on and discusses some of the main issues regarding AI and its impact on humanity in the near future. He considers the short-term effects of the development of AI such as space exploration, laws, AI weapons, jobs and wages, and the quest for human-level intelligence or AGI (Artificial General Intelligence). He often cites examples like *IBM Deep Blue*, *IBM Watson*, *Google DeepMind* (computer programs that can beat humans in chess, *Jeopardy* and *Go*) as well as self-driving cars, financial software and computer games which, in my opinion, brings the topic closer to the reader. I found this chapter very interesting and practical because the reader can begin to understand why securing AI is not only critical for the near future, but also how an inadequate security of AI could lead to catastrophic consequences in a much farther period of time in terms of public safety, financial stability, transportation, energy, healthcare, space exploration, and so on.

After discussing the current progress and possible issues in AI, Tegmark takes the reader on a journey through the "intelligence explosion" that will happen if one day we succeed in building human-level AGI, referring to the scenario of Prometheus and the Omega Team overtaking the world.

The fifth chapter includes a broad range of very interesting possible scenarios and consequences that could occur between intelligent machines and humans in a more distant future (*The Next 10,000 Years*), all depending on how we design AI's path, and whether the superintelligence will stay on those paths or decide to take a path for itself. Tegmark describes both positive and negative relations and potential outcomes, from a peaceful coexistence of humans and machines to the enslavement of machines and even to the complete overtaking of machines and extinction of humanity. This chapter can be very interesting for philosophers because it deals with concepts of political philosophy such as libertarianism, totalitarism, egalitarianism, Orwellianism, freedom, social structures, political power, property rights and so on, all in a relationship of integration in a society between humans and intelligent machines.

The sixth chapter is a speculation about life's future potential aided by technology and how could it flourish in the next billion years and beyond, not only in our Solar System but in all the possible cosmos. Tegmark describes the various ways the superintelligence could develop, whether it would become a rogue, how humans would interact with it, and would it prevent the predicted end of our universe. It takes a physicist to imagine how far life could life progress if limited only by the laws of physics. This part of the book is pretty astonishing even if most of it can hardly be achievable due to various limitations and possible cosmic wipeout. There is a lot of futurology in Tegmark's book which can be a little bit frustrating, especially about the things that we are highly unlikely to be able to predict, though at least the author recognizes this and points it out.

The remaining chapters explore concepts in physics, goals, ethics, the subject of consciousness and meaning, and then investigate what society can do to help create a desirable future for humanity. Tegmark believes that, in the future, when we create intelligent machines we could consider them, in some sense, as our descendants; we would be very proud of what they can do, they would have our values, and would do all the great things that we couldn't do. Even if they choose to eliminate us, they will live on and continue the story of life in our part of the observable universe. But what if those machines are zombies without any consciousness? Then if we humans eventually go extinct there will be nobody experiencing anything. It's like our whole universe had died for all intents and purposes. Tegmark believes that it's not our universe giving meaning to us, but we as conscious beings are giving meaning to our universe. The meaning comes from our experience. If there's nobody experiencing anything, our whole cosmos just goes back to being a giant waste of space. For these various reasons is very important to understand what it is about information processing that gives rise to what we call consciousness. Tegmark discusses what consciousness could be, saying that "consciousness is the way information feels when being processed in certain ways" and speculates that it must be substrate-independent, similarly to remembering, computing and learning (474–475). Tegmark argues that the risks of AI come not from malevolence or conscious behavior intrinsically, but rather from the misalignment of the goals of AI with those of humans. In Tegmark's words, "the real risk with artificial general intelligence isn't malice but competence. A superintelligent AI will be extremely good at accomplishing its goals, and if those goals aren't aligned with ours, we're in trouble" (407). Still, there are a lot of questions that humans should try to answer before any superintelligence is created, says Tegmark. He finishes his book optimistically, describing his work at the Future of Life Institute he has founded, which aims to ensure that we develop not only technology, but also the wisdom required to use it beneficially.

There is no doubt that the progress of AI can become an issue that needs thinking, writing and discussing about, and I believe Tegmark did a great job with *Life 3.0.* The book probably needs to be read alongside Nick Bostrom's *Superintelligence: Paths, Dangers, Strategies* and other recent books in the field to get a full picture. My opinion is that it would be even more effective if it was a bit shorter. Some chapters are intensely exciting and informative, others, including *Our Cosmic Endowment: The Next Billion Years and Beyond,* are a little bit too long and pretentious. Some chapters feel like fillers, put there just to make the book thicker, they add little to no useful information on AI. On the other hand, the long awaited chapters on the ethical questions and consciousness, which would have made the book more interesting for me, especially from the perspective of a philosopher, are just a scratch on the surface and do not delve into the depth of these issues.

Nevertheless, *Life 3.0: Being Human in the Age of Artificial Intelligence* by Max Tegmark is a great book that I'd recommend to anyone interested in the topic of AI, the long-term effects of future technology and its ramifications on all aspects of mankind. People working in the AI should definitely read this book so that they understand the broader concerns surrounding this area. I intend to read it again as the discussion on AI will get more and more interesting and gain importance over time.

Personally, this topic is very close to me. During the last couple of years I took a deep interest in AI research, consciousness and the possibility of creating human-level AGI and *Superintelligence*. Last year, while I was finishing my master thesis on the philosophical problems of AI, Max Tegmark's book was published and instantly hit the bestseller lists in September and October 2017. Too bad it wasn't published earlier because it would be of great help for me, it's the kind of book I was looking for at that time, simple, well-rounded and up to date with the recent events in AI.

I believe that *Life 3.0: Being Human in the Age of Artificial Intelligence* is a very accessible and highly readable book even for readers with no background knowledge in the field of AI. Due to Tegmark's simple style of writing and avoiding fancy words, he successfully gave clarity to the many faces of AI, starting from the history of the field to the implications of recent accomplishments in AI and the more detailed analysis of how we might get from where we are today to human level AGI or even to *Superintelligence,* a general intelligence far beyond the human level. The author talks about every possible argument and every point of view regarding AI that it's hard to find the main conclusion, but he presents multiple viewpoints which gives the reader a well-rounded perspective to come to his own conclusions. Max Tegmark is an interesting and provocative thinker; he uses stories that seem like SF novels to show the possible ways that AI could develop. He did an amazing job explaining the most likely outcomes in a simple manner that even readers lacking technology knowledge could understand it. With the description of an AI evolution closer than we imagine it, he enables the reader to look its possibility, pros and cons, as well as its impact on humanity (jobs, laws, weapons) with a perspective of its future potential.

This book could be seen as a challenge for humans interested in the future of life, intelligence and consciousness, a challenge on how to create a benevolent future civilization of humans merged with a possibly even greater intelligence than our own. I truly believe that this will be the most important conversation of our time and we should ask ourselves what we can do to improve our future coexistence with AI and avoid the risks that might get us in trouble.

IVAN SAFTIĆ
University of Rijeka, Rijeka, Croatia

Table of Contents of Vol. XVIII

Articles

Book Discussions

Book Reviews

Croatian Journal of Philosophy is published three times a year.
It publishes original scientific papers in the field of philosophy.

Croatian Journal of Philosophy is indexed in
The Philosopher's Index, PhilPapers, Scopus, ERIH PLUS
and in *Arts & Humanities Citation Index* (*Web of Science*).

Payment may be made by bank transfer

SWIFT PBZGHR2X
IBAN HR4723400091100096268

Croatian Journal of Philosophy is published with the support
of the Ministry of Science and Education
of the Republic of Croatia.

Instructions for Contributors

All submissions should be sent to the e-mail: cjp@ifzg.hr. Submissions must be in English and formatted to be double-spaced with suitably wide margins, an A4 page size, and automatic page numbering.

Articles are normally no longer than 8,000 words of main text including bibliography. The Journal will consider longer papers, but, once these limits are exceeded, authors should bear in mind the editorial policy that the acceptance bar raises with increasing length.

Manuscripts should be compiled in the following order: cover page; title; abstract (not exceeding 200 words); keywords (3 to 6); main text; appendices (as appropriate); references.

All the authors of a paper should include their full names, affiliations, postal addresses, telephone and fax numbers and email addresses on the cover page of the manuscript. If a paper is co-written, one author should be identified as the Corresponding Author. The cover page must be submitted as a separate document. All submitted manuscripts must be prepared for blind review, with revealing acknowledgements and self-identifying references removed.

Sources are cited in the text by the author's last name, the publication date of the work cited, and a page number if needed, e.g. (Barber 2007: 324). Full details appear in the reference list in which the year of publication appears immediately after the author's name:

Barber, A. 2007. "Linguistic Structure and the Brain." *Croatian Journal of Philosophy* 21 (7): 317–341.

Williamson, T. 2013. *Identity and Discrimination*. Oxford: Wiley-Blackwell.

The publication of a manuscript in the Croatian Journal of Philosophy is expected to follow standards of ethical behavior for all parties involved in the publishing process: authors, editors, and reviewers. The journal follows the principles of the Committee on Publication Ethics (https://publicationethics.org/resources/flowcharts).

ISSN 1333-1108

9 771333 110001