Vol. V, No. 37, November 2013

# Disputatio

International Journal of Philosophy

## Articles

## Critical notices

## Book Symposium on Berit Brogaard's *Transient Truths* (Oxford University Press 2012)

## Book reviews

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

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 Articles of a purely exegetical or historical character will not be published.

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

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

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

All material published in *Disputatio* is fully copyrighted. It may be printed or photocopied for private or classroom purposes, but it may not be published elsewhere without the author's and *Disputatio*'s written permission. The authors own copyright of articles, book reviews and critical notices. *Disputatio* owns other materials. If in doubt, please contact *Disputatio* or the authors.

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

| PUBLISHED BY | SPONSORS |
|---|---|
| VNIVERSITAS<br>Centro de Filosofia<br>OLISIPONENSIS<br>**Centro de Filosofia**<br>Universidade de Lisboa | **FCT**<br>**Fundação para a Ciência e a Tecnologia**<br>MINISTÉRIO DA CIÊNCIA E DA TECNOLOGIA<br>Apoio do Programa Operacional Ciência, Tecnologia,<br>Inovação do Quadro Comunitário de Apoio III. |

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

# Disputatio

## International Journal of Philosophy

Vol. V, No. 37, November 2013

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

*Disputatio* Lecture 2012

# Scepticism and Implicit Bias[1]

**Jennifer Saul**
University of Sheffield

BIBLID [0873-626X (2013) 37; pp. 243-263]

The goal of this paper is to explore the idea that what we know about implicit bias gives rise to something *akin to* a new form of scepticism. I am not wedded to the idea that the phenomenon I am pointing to should be called 'scepticism', but I am convinced that it is illuminating to examine the ways in which it does and does not resemble philosophical scepticism. I will call what I am discussing 'bias-related doubt'.

In some ways, bias-related doubt is stronger than traditional forms of scepticism, while in others it is weaker. In brief: I will be arguing that what we know about implicit biases shows us that we have very good reason to believe that we cannot properly trust our knowledge-seeking faculties. This does not mean that we might be mistaken *about everything*, or even everything in the external world (so it is weaker than traditional scepticism). But it does mean that we have *good reason* to believe that we are mistaken about a great deal (so it is stronger than traditional forms of scepticism). A further way in which bias-related doubt is stronger than traditional scepticism: this is doubt that demands action. With traditional scepticism, we feel perfectly fine about setting aside the doubts we have felt when we leave the philosophy seminar room. But with bias-related doubt, we don't feel fine about this at all. We feel a need to *do something* to improve our epistemic situation. Fortunately, though, it turns out that there is much we can do. However, much of what needs to be done cannot be done on a purely individual basis.  So although scepticism has sometimes been treated by feminists as a paradigmatic case of the

[1] This Disputatio Lecture was delivered at the 5th Meeting of the Portuguese Society for Analytic Philosophy, which took place in Braga at the University of Minho, on September 14th 2012.

*Disputatio*, Vol. V, No. 37, November 2013

Received: 23/10/2013 Accepted: 23/10/2013

excesses of individualist philosophy[2], this form of scepticism cannot be fully responded to individualistically.

## 1 Implicit biases

There is a vast and still-growing literature on implicit bias, which I'll only be dipping into here. Very broadly speaking, these are largely unconscious tendencies to automatically associate concepts with one another.[3] Put like this, they don't sound very interesting or worrying. But the ones on which attention by philosophers has focused are both very interesting and very worrying. These are unconscious, automatic tendencies to associate certain traits with members of particular social groups, in ways that lead to some very disturbing errors: we tend to judge members of stigmatized groups more negatively, in a whole host of ways. Rather than attempt a general overview, I will give examples of the sorts of errors that will be our concern here.

### Curriculum vitae

CV studies take a common, and beautifully simple form. The experimenters ask subjects to rate what is in fact the same CV, varying whatever trait they want to study by (usually) varying the name at the top of it. When they do this, they find that the same CV is considered much better when it has a typically white rather than typically black name, a typically Swedish rather than typically Arab name, a typically male rather than typically female name, and so on. The right name makes the reader rate one as more likely to be interviewed, more likely to be hired, likely to be offered more money, and a better prospect for mentoring. These judgments are very clearly being affected by something that *should* be irrelevant — the social category of the person whose CV is being read. Moreover, the person making these mistaken judgments is surely unaware of the role that social category is playing in the formation of their views of the candidates.

[2] See, for example, Scheman 2002.

[3] For a great deal more precision about the many different ways of characterizing implicit bias, and the many sorts of implicit biases there are, see Holroyd and Sweetman (forthcoming).

Significantly, the most recent of these studies (Moss-Racusin 2012), on the evaluation of women's CVs, showed that women were just as likely to make these problematic judgments as men. It also showed that these problems are not confined to an older generation: the tendencies were equally strong in all age groups.[4]

## Prestige bias

In a now-classic study, psychologists Peters and Ceci (1982) sent previously published papers to the top psychology journals that had published them, but with false names and non-prestigious affiliations. Only 8% detected that the papers had already been submitted, and 89% were rejected, citing serious methodological errors (and not the one they should have cited — plagiarism). This makes it clear that institutional affiliation has a dramatic effect on the judgments made by reviewers (either positively, negatively, or both). These are experts in their field, making judgments about their area of expertise — psychological methodology — and yet they are making dramatically different judgments depending on the social group to which authors belong (member of prestigious VS non-prestigious psychology department).

## Perception

Studies of so-called 'shooter bias' show us that implicit bias can even influence perception. In these studies, it has been shown that the very same ambiguous object is far more likely to be perceived as a gun when held by a young black man and something innocent (like a phone) when held like by a young white man.[5] (The same effect has been shown with men who appear Muslim versus men who appear non-Muslim (Unkelbach et al. 2008). In some of these experiments, the subjects' task is to shoot in a video game if and only if they see an image of a person carrying a gun. Subjects' 'shooting' is just as

[4] See, for example, Bertrand and Mullainathan 2004; Rooth 2007; Moss-Racusin et al. 2012; Steinpreis et al. 1999.

[5] See, for example Correll et. al. 2002, 2007; Greenwald, Oakes, & Hoffman, 2003; Payne, 2001; Plant & Peruche, 2005.

you'd expect given their perceptions. These show that implicit bias is getting to us even before we get to the point of reflecting upon the world — it affects our very perceptions of that world, again in worrying ways.[6]

## Moral and political consequences

Now let's explore some consequences of this. First, there are some obvious morally and politically significant consequences. We are very likely to make inaccurate judgments about who is the best candidate for a job, if some of the top candidates are known to be from stigmatised groups. We are very likely to mark inaccurately, if social group membership is known to us and the group we are marking is not socially homogeneous. We are very likely to make inaccurate judgments about which papers deserve to be published, if social group membership is known to us. We may both over-rate members of some groups and under-rate others. Worse yet, we are misperceiving harmless objects as dangerous, and potentially acting on this in truly appalling ways. All of this *should* be tremendously disturbing to us. It means that we are being dramatically *unfair* in our judgments, even though we are doing so unintentionally. We are treating members of stigmatised groups badly, even if we desperately desire to treat them well. Moreover, what we are doing will help to ensure that this unfair treatment is continued: the results of these decisions will help to maintain the stereotypes that currently exist, which cause members of stigmatised groups to be treated unfairly. 'Vicious circle' seems a particularly apt phrase to describe the situation.

## Epistemological consequences

But I want to focus now on some epistemological aspects of this situation. First, some relatively obvious ones, starting from those within philosophy. The unfairness described above means that there are almost certain to be some excellent students receiving lower marks and less encouragement than they should; some excellent philosophers not getting the jobs they should get; and where anonymous

[6] For much more on how perception is affected, see Siegel 2013.

refereeing and editing is not practised, there is some excellent work not being published. Philosophy as a field is the worse for this: it is not as good as it could, or should, be. (For more on this, see Beebee and Saul 2011, Saul forthcoming.) Obviously, much the same will go on in other areas of academia, especially those that are as male-dominated as philosophy. Outside philosophy, there are similar effects, as the testimony of members of stigmatised groups is taken less seriously than it ought to be (Fricker 2007). Their views are less respected, and they are given less of an opportunity to participate fully in discussions and decision-making. As Chris Hookway (2010) has noted, a particular problem may lie in their *questions* not being taken seriously.

Now, some less obvious epistemological aspects of the situation, again focussing on philosophy. When we misjudge a paper's quality, we're making a mistake about the quality of an argument.[7] Moreover, our evaluation of that argument is being influenced by factors totally irrelevant to its quality: it's being influenced by our knowledge of the social group of its author. Worse yet, this influence operates below the level of consciousness — it's unavailable to inspection and rational evaluation. This means we may be accepting arguments we should not accept and rejecting arguments we should not reject. Many of our philosophical beliefs — those beliefs we take to have been arrived at through the most careful exercise of reason — are likely to be wrong.[8]

But now a cynical objection emerges, and here's how it goes: philosophy *is* in fact incredibly homogeneous (only 17% of those employed full-time as philosophers in America are women[9]). When we're deciding which argument to accept, we're mostly deciding which argument from a white, cis-gendered, middle-class, able-bod-

[7] Here I am assuming that philosophers will be prone to the same sorts of errors as others. They have not actually been studied.

[8] I am *not* saying that we are affected only by biases. Of course, a part of what we are doing is applying our skill in evaluating philosophy, and sometimes we will get things right. My claim is just that these judgments will often be distorted, to a variable extent, by biases.

[9] See <http://kathrynjnorlock.blogspot.co.uk/p/my-apa-csw-report-on-women-in.html>.

ied man to accept. So, while we may have wrongly rejected an argument presented with a working class accent, a dark skin or a woman's name, surely this won't have happened very often. Our philosophical beliefs are largely safe, if only because philosophy is so disturbingly homogeneous.

But there are two responses to the cynical objection. The first is to recall the existence and force of prestige bias. In prestige bias, the stigmatised group is people who are less prestigious — their names are less famous, their institutions less notable, their accomplishments (so far) less impressive. And the advantaged group is those who benefit from prestige — those with famous names, highly ranked institutions, and accomplished careers. We are extremely likely to be making errors about people from both of these groups. Since it is surely the case that we encounter plenty of people from both these groups in e.g. hiring and refereeing papers, there are many opportunities for us to be affected by these biases. The second response is to note that while e.g. 17% is a lamentably small proportion of the profession, it is not so small that we're unlikely to encounter a woman in philosophy. We have, in fact, ample opportunities to misjudge the quality of work from members of at least some under-represented groups. And this is enough to give bite to bias-related doubt. So, we are likely to be making errors in our evaluation of philosophical arguments.

It is important to see that this is not *just* a matter of what Miranda Fricker (2007) has called testimonial injustice. Fricker argues that the social group to which a person belongs will often have a dramatic effect on our willingness to treat them as a credible source of knowledge. We will be less likely to accept the testimony of those from stigmatised groups. One thing implicit bias adds to this picture is just a matter of scale: research shows these problems to be far more widespread than would otherwise be apparent. But another, even more important addition, is that implicit bias doesn't just affect our judgments of people's *credibility* when deciding whether to accept their testimony or not. Mistaken as our credibility judgments are, at least we know that these are judgments about who to take seriously. We recognise that we are making judgments about people, and this is what we mean to be doing. The research on implicit bias shows us that we are actually being affected by biases about social groups *when we think we are evaluating evidence or methodology.* When considering testimony, it makes sense that we need to make judgments about how credible an individual is. But when psychologists assess the methodology of a study — or when philosophers assess the quality of an argument — they shouldn't be looking at the credibility of an individual at all. They should be looking just at the study, or the argument. And yet when implicit bias is at work, we are likely to be affected by the social group of the person presenting evidence or an argument even when were are trying to evaluate that evidence or argument itself. Implicit bias is not just affecting who we trust — it's affecting us when we think we're making judgments that have nothing to do with trust. It's leading us into errors based on social category membership when we think we're making judgments of scientific or argumentative merit.

But why should that unsettle us? We know already that most of what is currently accepted as science is likely to be proven false within centuries, and possibly decades. But notice: my claim is not that we're likely to be accepting some falsehoods, or even a lot of falsehoods. That's not unsettling. My claim is that we're likely to be *making errors*. Moreover, we're likely to be making errors of a very specific sort. It's *not* that we're likely to get some really difficult technical bits wrong, or that we're likely to get things wrong if we're really exhausted, or drunk. It's that we're likely to let the social identity of the person making an argument affect our evaluation of that argument. It is part of our self-understanding as rational enquirers that we will make certain sorts of mistakes. But not this sort of mistake. These mistakes are ones in which something that we actively think *should not* affect us does.

Worse yet, our errors are not confined to the professional arena, or to what we take to be carefully thought-out judgments about the quality of arguments that we encounter. The studies of shooter bias show us that as humans in the world, we are making errors in *perception* due to implicit bias. The very data from which we begin in thinking about the world — our perceptions — cannot be relied upon to be free of bias. Once more, this is clearly well beyond the worries raised by testimonial injustice.

The best way to see why these mistakes are — and should be — so unsettling to us as enquirers is to compare the situation of one who learns about implicit biases to the situations of people considering various sorts of sceptical scenarios.

# 2 Comparison to sceptical scenarios

## 2.1 Comparison to traditional scepticism

In a traditional skeptical scenario, we are confronted with a possibility that we can't rule out — that we're brains in vats, or that tomorrow gravity might not work any more. Considering this scenario is meant to make us worry that we don't know (many of) the things that we take ourselves to know, or that we are unjustified in having (many of) the beliefs that we do. And a standard response is that these should worries not grip us, because we have no reason at all to suppose that these possibilities obtain. Doubt induced by implicit bias is unlike this: we have *very good reason* to suppose that we are systematically making errors caused by our unconscious biases related to social categories. In this way, then, the doubt provoked by implicit bias is stronger than that caused by considering skeptical arguments.

But, one might think, it's not really all that troubling. The doubt caused by implicit bias, surely, is a localized one. It seems, at first, to be like the sort of doubt we experience when we discover how poor we are at probabilistic reasoning. We have extremely good reason to think we're making errors when we make judgments of likelihood. But this sort of doubt doesn't trouble us all that much because we know exactly when we should worry and what we should do about it: if we find ourselves estimating likelihood, we should mistrust our instincts and either follow mechanical procedures we've learned or consult an expert (if not in person, then on the internet). This kind of worry is one that everyone can accept without feeling drawn into anything like skepticism. And it may seem at first that bias-related doubt is like this.

The problem starts to become vivid when we ask ourselves *when* we should be worried about implicit bias influencing our judgments. The answer is that we should be worried about it whenever we consider a claim, an argument, a suggestion, a question, etc from a person whose apparent social group we're in a position to recognize. Whenever that's the case, there will be room for our unconscious biases to perniciously affect us. Most discussed in the literature so far (see Fricker 2007), we might make a mistaken judgment of credibility when assessing testimony. But we also might fail to listen properly to a contribution; fail to carefully consider a question; judge an argument to be less compelling or original than it is; think the evidence presented is worse than it is. And, importantly, we can be adversely affected in a positive direction as well. When assessing a contribution from someone who are biases favour, we may grant more credibility than their testimony deserves; we may think their arguments are better than they are, perhaps failing to notice flaws that we would have noticed if the arguments were presented by someone else; we may take their evidence to be better than it is, and so on.

And *this* is going to happen a great deal. It happens whenever we are dealing with the social world in a non-anonymised manner. Since the world is only rarely anonymised for us, this will happen nearly all the time. Much of our knowledge comes from testimony, or from arguments or evidence that we are presented with. Those testifying, or presenting the arguments or evidence, are usually people. And people are generally (though not always) perceived by us as members of social groups. Moreover, much of the knowledge we already have has come to us in this way. Our acceptance or rejection of testimony, arguments, evidence and the like has shaped the worldviews we have now. And this acceptance or rejection was, we can be fairly certain, distorted by the perceived social groups[10] of those presenting the testimony, arguments or evidence. Worse yet, we cannot even go back and attempt to consider or correct errors that we might have made — we are very unlikely to remember the sources of these beliefs of ours.

Chris Hookway has argued, compellingly, that the most effective and interesting construal of skeptical challenges is to see them as challenges to our ability to enquire responsibly. A skeptical argument is a challenge to the reliability of what he calls ‘our cognitive instruments’, and it demands that we demonstrate that we are not being irresponsible in relying on our ordinary belief-forming meth-

[10] I phrase it this way because what affects us as audiences is what social group we *take* the speaker to be a member of, not what social group they are actually a member of.

ods. Hookway considers two kinds of challenges to our belief-forming methods.

The first kind of challenge, very localized and not at all threatening, is something like the worry we've seen about our probability judgments. Here we're given very good reason to distrust our cognitive instruments, but only for certain tasks. We can easily demonstrate that we're being responsible, as long as we take special measures regarding probabilitistic judgments, and that's a pretty straightforward thing to do. It's kind of like the discovery (which most of us don't remember making) that we can't really see very well in the dark: we turn on a light when we encounter darkness, and the problem's solved. It doesn't worry us the rest of the time. So this kind of challenge needn't worry us too much.

The next kind of challenge is broader — if we were to become convinced that we *really did need to worry* about the possibility of being brains in vats, it would lead us to question almost everything we think we know. And there would be no easy solution. But, following Peirce, Hookway maintains that this isn't a real doubt: we don't really doubt the existence of the external world, and 'we should not doubt in philosophy what we do not doubt in our hearts' (2002: 248). Both Hookway and Peirce assign real philosophical weight to this failure to doubt, taking it to obviate the need to demonstrate that we are not being irresponsible. (Though both also accept the fallibility of this.)

Bias-related doubt is different from this, though. It is very much a real doubt. Interestingly, this is because it combines one feature each from the two not-so-worrying forms of doubt just considered.

1. We have been given good reason to think that we are very likely to be making the errors it points to. This makes the doubt genuinely compelling, and we feel at as genuinely compelling. In this way, it is like the worry about our probability judgments.
2. It is broad in its scope. It arises with regard to any beliefs that might have been unconsciously shaped by our implicit attitudes about members of social categories, and these are an enormous number of our beliefs. Moreover, we don't in general even know which beliefs these are. This gives it the

> kind of breadth that leads to a much greater worry than the very containable concerns about probabilistic reasoning.

Hookway writes that there are three key features to 'an interesting skeptical challenge'. (1990: 164)

1. It must make reference to 'part of our practice of obtaining information about our surroundings which we find natural, which it does not ordinarily occur to us to challenge.'
2. '[I]t must have a certain generality: challenges to the reliability of particular thermometers may lead us to lose confidence in that particular instrument; they do not lead us to lose confidence in ourselves as inquirers.'
3. '[I]t must intimate that the feature of our practice which it draws attention to *could not* be defended.'

It seems to me that bias-related doubt easily meets each of these criteria. The practices called into question ones that we normally don't think to question: our 'instinctive' sense that someone is credible, that a reason is convincing, or that an argument is compelling. There is definitely generality — this isn't like challenges just to probabilistic reasoning, which Hookway rightly flags as not that worrying because those challenges are very contained. Instead, it's challenges to the ordinary ways that we assess reasons, arguments evidence and testimony. Finally, the feature it calls attention to — our judgments are illicitly influenced by irrelevant matters in a way that frequently leads to injustice — is deeply indefensible.

What the literature on implicit bias shows us is that we *really should not* trust ourselves as inquirers. As Hookway argues (2003: 200), 'we can persevere with our inquiries only if we are confident that… our reflection will take appropriate routes'. But we have now discovered that our reflection takes wholly inappropriate routes: we are not only failing to assess claims or arguments by methods that we endorse but we are instead assessing them by methods that we actively oppose. As he notes, only a part of the process of deliberation is conscious, and we need to be able to trust the habits of thought that underpin the unconscious bits. (Hookway 1990: 11) we need to trust not just that they will guide us to truth but that they are based in values that we consider our own. Hookway raises the values concern when discussing an obsessive who is unable to stop repeatedly rehearsing doubts that he does not fully endorse, but the concern arises even more strongly in the case of biases against members of stigmatized groups. The literature on implicit bias shows us not just that our habits can't be relied on to lead us to truth, but also that — insofar as they can be described as based in values at all — they are likely to be based in values that we (most of us, anyway) find repugnant. It is difficult to see how we could ever properly trust these again once we have reflected on implicit bias. And, Hookway (2000: Chapter 10) argues, self-trust is a necessary condition of responsible inquiry.

## 2.2 Comparison to live sceptical scenarios

Bryan Frances's work on 'live sceptical scenarios' (Frances 2005), provides another instructive comparison. Frances characterizes traditional skeptical arguments as relying on the fact that certain hypotheses that cannot be ruled out. He notes that responses to these often involve pointing out that, while these hypotheses cannot be ruled out, they are nonetheless not really *live* — they are so implausible that we can't really take them seriously. His book is devoted to arguing that there are skeptical hypotheses that are not like this. In his live skeptical scenarios, 'there are compelling scientific and philosophical reasons to think that the hypotheses are actually true'. Therefore, the traditional replies do not apply.

Now this looks quite a lot like what I have called Bias-Related Doubt. The hypotheses are ones for which there is compelling reason for thinking that they are true. But on closer inspection, it turns out that these reasons are far less compelling. The hypotheses in question are things like eliminativism about belief and error theory about colour. And the reasons for thinking that they are still live is that some sensible people who know a great deal endorse (or might endorse) these theories on the grounds of compelling scientific or philosophical reasons. But this falls a good deal short of what I have argued about bias-related doubt. Here the hypothesis is that we are frequently making errors that have their root in implicit bias. My claim is not just that the hypothesis is live — that sensible and knowledgeable people might endorse it on the basis of good reasons. Instead, it's that *we all have very good reason to believe that it is true*. And this is much stronger than the claim that a hypothesis is live. We will see that there are also differences with regard to how we should respond.

## 3 What should we do?

The skepticism created by learning about implicit bias differs dramatically from most other forms of skepticism in that it leads to the conclusion that we should change our behaviour. A striking feature of the sorts of skepticism that have tended to dominate discussion in recent times is that *even if* we became convinced by them, we would not feel the need to change in anything about our behaviour: accepting that I don't know whether I'm a brain in a vat or not simply doesn't affect how I will go about living my life. Becoming a sceptic of the traditional sort doesn't lead me to decide differently about anything in the course of my every day life, or to alter my behaviour in any way.

But not all forms of skepticism are like these in their lack of impact on behaviour: Pyrrhonian skepticism was meant to have a large and salutary impact on one's life. The convinced Pyrrhonian sceptic would learn to simply accept appearances rather than striving for belief.

> 'If he avoids 'belief', the Pyrrhonist 'acquiesces in appearances': he is guided by sensory appearances and by bodily needs and natural desires; he conforms to the prevailing customs and standards of his society.'[11]

Accepting appearances and conforming to prevailing customs and standards, of course, is very much *not* what a would-be responsible enquirer should feel moved to do after learning about implicit bias. For the literature on implicit bias shows that the way things appear to us is perniciously affected by biases that we are unaware of and would repudiate if we became aware of them. To put it bluntly, accepting appearances would mean acquiescing in one's reaction of fear at the sight of a black man; and acquiescing in one's greater sense of approval when looking at a CV with a man's name at the top of it.

[11] Hookway (1990: 6).

That these would not rise to the level of belief may mean that we're not committed to falsehoods. But the behaviours we would be led to would be just as troubling. As Hookway notes (1990: 18), the Pyrrhonist's 'is a very conservative outlook: the appearances he relies on are salient for him because of their conventional role.' Relying on the conventions of one's society is deeply cast into doubt by the literature on implicit bias.

The skepticism produced by implicit bias demands action. There are several reasons for this. The first reason is that the skeptical scenario is one that is troubling in a very different way from more traditional skeptical scenarios. If you actually are a brain in a vat, you're probably doing about as well with your life as you can. It's not clear that you would make different choices if you knew the scenario to hold. (And this is just as true for the live skeptical scenarios Frances considers, like those based in eliminativism or colour error theory.) But if you actually are basing lots of decisions on the social categories that people you encounter belong to, then you're clearly not doing as well as you can. You're making the wrong decisions epistemically speaking: taking an argument to be better than it is, perhaps; or wrongly discounting the view of someone you should listen to. You're also making the wrong decisions practically speaking: assigning the wrong mark to an essay, or rejecting a paper that you should accept. Finally, you're making the wrong decisions morally speaking: you are treating people unfairly; and you are basing your decisions on stereotypes that you find morally repugnant. So when the possibility is raised that you're doing this, it should not be possible to shrug it off in the way that it's perfectly reasonable to shrug off the brain in a vat possibility. Worse yet, it's not just the *possibility* that's raised: the research on implicit bias suggests that it's very likely that you're doing these things, with respect to at least some social categories.

But usually, you can't do anything at all to rule out the skeptical scenarios. And the same is true when it comes to any particular instance of the implicit bias skeptical scenario. Did I judge that woman's work to be less good than it was due to her gender? I will never know, because I won't get the opportunity to assess it without knowledge of her gender. And the same is true for certain more general versions: have I based much of what I think I know on epistemically irrelevant factors like social categories? I'm not going to be able to find out. So is there *anything* one can do? Not for past cases like these. However, I can act so as to reduce the likelihood of this happening in future instances.

Importantly, though, some of the most obvious things to do just don't work. Getting a woman to judge another woman's work is a poor check against bias, since both men and women are likely to hold biases causing them to negatively judge women's work (recall Moss-Racusin's 2012 CV study). Trying hard to be unprejudiced can backfire, if one doesn't go about it in just the right way (Legault et al. 2011). Reflecting on past instances in which one managed to do the right thing makes one *more* likely, not less likely to be biased (Moskowitz and Li). So what should one do?

Fortunately, there are some things we can do. Obviously anonymising can prevent us from even being aware of the social group that might trigger our implicit biases.[12] But anonymising is not a solution that's always available or appropriate, so it's fortunate that psychologists are discovering a lot of surprising interventions that seem to reduce the influence of implicit biases. We can spend time thinking about counter-stereotypical exemplars (members of stereotyped groups who don't fit the group stereotypes)[13]. We can carefully form implementation intentions — not 'I will not be influenced by race' but 'when I see a black face I will think 'safe'' (Stewart and Payne 2008). We can spend a few hours engaging in Kawakami's negation training, in which we practice strongly negating stereotypes (Kawakami et al. 2000). But this might not work, unless we use Johnson's (2009) variant in which we think 'NO, THAT'S WRONG!' while pressing a space bar whenever presented with a stereotypical pairing. We can reflect on past instances in which we *failed* in efforts to be unbiased, thereby activating our motivation to

[12] This worked beautifully with orchestras, which began holding auditions behind screens, dramatically increasing their percentages of female members. And it is now standard practice in the UK to mark students' work anonymously, which is supported by the Union of Students for just this reason: <http://www.nusconnect.org.uk/campaigns/highereducation/archived/learning-and-teaching-hub/anonymous-marking/>. For research on anonymous marking see Bradley 1984, 1993.

[13] Blair 2002, Kang and Banaji 2006.

control prejudice (Moskowitz and Li). And these are just a few examples.

Interestingly, some very effective interventions — like Kawakami's negation training — are widely viewed as far too demanding for widespread adoption. Alex Madva (manuscript), however, has argued extremely compellingly that these have been dismissed far too quickly. And he has a point — what's a few hours of slightly tedious exercises if it can actually make me less prejudiced? The arguments I have presented here suggest that we may well also have very strong *epistemic* reasons as well for adopting these techniques. If we don't try to overcome the pernicious influences of these biases, we are not being responsible enquirers.[14]

Importantly, though, we are unlikely to completely eliminate the threat of error. Implicit bias could be affecting one's reasoning at almost any point — it is very hard to judge when social group membership is having a pernicious influence. So it is much trickier to correct for than other factors that are known to make one unreliable (e.g. 'don't make important decisions when drunk'). If we knew that we were about to enter a situation in which implicit biases might impair our thinking, and we knew exactly which biases would be relevant, we could formulate appropriate implementation intentions, like 'If I see black person, I will think 'safe''. But we don't in general know which stigmatized social groups we will encounter at which points, or what stereotype will be relevant. (Thinking 'safe' when we see a black person will not help us to more accurately assess the quality of their written work.) Moreover, we don't know what sorts of cognitive task might be relevant. So far, I have focused mostly on assessments of quality of argument, or of believability. But implicit biases surely affect other epistemically relevant matters as well: they might lead me to ask the wrong questions, or to neglect the right ones. Implementation intentions are a powerful device for controlling the expression of biases, but by their nature they target very specific behaviours. They cannot provide the general sort of reshaping of the cognitive faculties that would be needed to fully combat the

[14] Madva also responds to criticisms that these techniques are not effective enough, and that they are too individualistic, focusing as they do on individual thinkers rather than societal reform.

influence of implicit biases. At the end of the paper, I'll discuss what this limitation to our individual corrective measures means for us.

## 4 Our rational capacities

Miranda Fricker is one of the few epistemologists who has thought long and hard about the negative epistemic effects of stereotypes. Her focus, however, is on the way that these affect evaluations of testimony from those that the stereotypes target, and she does not discuss the literature on implicit bias. This literature (as we have seen) shows the pernicious epistemic influence of stereotypes to extend far beyond evaluation of testimony. Still, Fricker's discussion is highly relevant: she argues that those who underrate the testimonies of others due to wrongful stereotyping of their social group are committing an injustice, and that they suffer from an epistemic vice. This terminology seems wholly appropriate to apply to those in the grip of pernicious implicit biases. It seems worth examining, then, what she says about correcting for prejudices.

Fricker suggests that there are two ways to be a virtuous agent in terms of accepting testimony. The first is to be 'naively' virtuous — to simply have credibility judgments that are not influenced by prejudice. She admits that this will be difficult to manage with respect to the prejudices of the culture/sub-culture one grows up in. The next is to reflectively correct one's judgments — to, for instance, think 'I'm white, and I may fail to give sufficient credibility judgment to black people as a result.' Or, alternatively, to notice that despite consciously believing women to be the equals of men, one tends to always take a man's word over a woman's. Noticing these things, she suggests, allows one to consciously raise the credibility one assigns to members of stigmatized groups. And this possibility, she suggests, is essential to our status as rational enquirers:

> 'The claim that testimonial sensibility is a capacity of reason crucially depends on its capacity to adapt in this way, for otherwise it would be little more than a dead-weight social conditioning that looked more like a threat to the justification of a hearer's responses than a source of that justification.' (84)

Extending this idea in a natural way, we would expect the capacity to consciously, critically, reflectively correct for one's biases quite generally to be crucial to one's epistemic capacities being capacities of reason.

Before we learn about implicit bias and what to do about it, it is genuinely unclear to me whether we have this ability to critically and reflectively correct for our bias. We could perhaps claim that we had the *ability* to do that (once we learned about the evidence, etc) but this claim would be so weak as not to amount to actually be very reassuring. Now, however, many of us do have the ability to critically and reflectively correct for our biases — at least once we have learned about their existence and studied the literature on what to do about them. Once we do that (and implement these techniques), we can responsibly claim that these capacities are not just dead-weight social conditioning. Importantly, though, this requires more than what Fricker imagined in her discussion: we are unlikely to notice through individualistic reflection the ways that our judgments are affected by social categories; and even we do notice this we are unlikely to hit upon the right strategies for fighting it. The only way that we can engage in the necessary sort of correction is not individualistically or introspectively, but by informing ourselves about what scientists have discovered about humans like ourselves. The correction is dependent not just on our rational faculties but on the deliverances of science.

In order to inquire responsibly, we must instead recognize that our epistemic capacities are prone to errors that we cannot learn about through first-person reflection; and that we must correct them using counter-intuitive mechanical techniques that draw not upon our rational agency but upon automatic and unconscious responses. We can consciously enlist these unconscious responses, and use them to improve our epistemic responses, but we cannot do this through rational and critical reflection alone.

Moreover, as I noted in the previous section, individual efforts are inevitably limited.

To fully combat the influence of implicit biases, what we really need to do is to re-shape our social world. The stereotypes underlying implicit biases can only fully be broken down by creating more integrated neighborhoods and workplaces; by having women, people of colour and disabled people in positions of power; by having men in nurturing roles; and so on. The only way to be fully freed from the grip of bias-related doubt is to create a social world where the stereotypes that now warp our judgments no longer hold sway over us. And the way to do this is to end the social regularities that feed and support these stereotypes. Can this be done? Who knows. It is a massive task — one whose importance and magnitude Elizabeth Anderson makes clear (for the case of race) in her *The Imperative of Integration.* But if it is not, we would seem to be stuck with bias-related doubt, and with the consequent lack of trust in our cognitive faculties. And this is in itself quite a fascinating result. Scepticism is generally thought of as a highly individualistic epistemic issue. It's about the would-be knower doubting the guidance of her own mind. But bias-related doubt shows us a social dimension to this. We have seen that the social world gives rise to a powerful form of doubt, and one that can only be fully answered by a sweeping and radical transformation of our social world.[15]

Jennifer Saul
Department of Philosophy
University of Sheffield
45 Victoria St
Sheffield S3 7QB
j.saul@sheffield.ac.uk

## *References*

Anderson, E. 2010. *The Imperative of Integration.* Princeton: Princeton University Press.

Bertrand, M., and Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? *American Economic Review*, 94, 991–1013.

Beebee, H. and Saul, J. 2011. *Women in Philosophy in the UK: A Report,* published by the British Philosophical Association and the Society for Women in Philosophy. (<9-08/Women%20in%20Philosophy%20in%20the%20UK%20(BPA-SWIPUK%20Report).pdf>)

Blair, I. 2002. The Malleability of Automatic Stereotypes and Prejudice. *Personality and Social Psychology Review*, 3: 242-261.

Bradley, C. 1984. Sex bias in the evaluation of students. *British Journal of Social*

[15] I had very useful discussions of this paper with several different audiences: The ENFA5 Conference in Braga, Portugal; the Eastern APA audience in Washington DC; and the departmental seminars at Nottingham and Southampton. I have also benefitted enormously from discussions with Louise Antony, Ray Drainville, Miranda Fricker, Teresa Marques and especially Chris Hookway — to whom this paper owes an obvious and enormous debt. (Though the errors are all mine.)

*Psychology*, 23: 2, 147-153.

Bradley, C. 1993 Sex bias in student assessment overlooked? *Assessment and Evaluation in Higher Education* 18: 1, 3-8.

Correll, J., Park, B., Judd, C., & Wittenbrink, B. 2002. The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, 83, 1314–1329.

Correll, J., Park, B., Judd, C., Wittenbrink, B., Sadler, M. S., & Keesee, T. 2007. Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, 92, 1006–1023.

Frances, B. 2005. *Scepticism Comes Alive,* Oxford: Oxford University Press.

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing.* Oxford: Oxford University Press.

Goldin, C. and Rouse, C. 2000. Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians. *The American Economic Review,* 90:4, 715-741.

Greenwald, A. G., Oakes, M. A. and Hoffman, H. 2003b. Targets of discrimination: Effects of race on responses to weapons holders. *Journal of Experimental Social Psychology*, 39, 399–405.

Holroyd, J. and Sweetman, J. Forthcoming. The Heterogeneity of Implicit Bias. In *Implicit Bias and Philosophy*, ed. by M. Brownstein and J. Saul. Oxford: Oxford University Press.

Hookway, C. 1990. *Scepticism,* London: Routledge.

Hookway, C. 2000. *Truth, Rationality, and Pragmatism: Themes From Peirce.* Oxford: Oxford University Press.

Hookway, C. 2003. How to be a Virtue Epistemologist. In *Intellectual Virtue: Perspectives from Ethics and Epistemology*, ed. by M. DePaul and L. Zagzebski. Oxford University Press.

Hookway, C. 2010. Some Varieties of Epistemic Injustice: Response to Fricker. *Episteme* 7:2, 151-163.

Johnson, I. R. 2009. *Just say 'No' (and mean it): Meaningful negation as a tool to modify automatic racial prejudice*. Doctoral dissertation, Ohio State University.

Kang, J. and Banaji, M. 2006. Fair Measures: A Behavioral Realist Revision of 'Affirmative action'. California Law Review 94: 1063-1118.

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S. and Russin, A. 2000. Just say no (to stereotyping): effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology* 78 , 871–888.

Legault, L., Gutsell, J., and Inzlicht, M. 2011. Ironic Effects of Antiprejudice Messages: How Motivational Interventions Can Reduce (But also Increase) Prejudice. *Psychological Science* 22(12) 1472–1477.

Madva, A. 2013. The Biases Against Debiasing. Paper presented at *Implicit Bias, Philosophy and Psychology Conference,* Sheffield April 2013.

Moskowitz, G. and Li, P. 2011. Egalitarian Goals Trigger Stereotype Inhibition: A Proactive Form of Stereotype Control, *Journal of Experimental Social Psychology* 47: 103–16.

Moss-Racusin, C., Dovidio, J., Brescoll, V., Graham, M., Handelsman, J. 2012. Science Faculty's Subtle Gender Biases Favor Male Students. *PNAS* 109 (41) 16395-16396.

Payne, B. K. 2001. Prejudice and perception: The role of automatic and controlling processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, 81, 181–192.

Peters, Douglas P., and Stephen J. Ceci. 1982. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences* 5:187–255.

Plant, E. A. and Peruche, B. M. 2005. The consequences of race for police officers' responses to criminal suspects. Psychological Science, 16, 180–183. Price-Waterhouse v. Hopkins, 109 S. Ct. 1775. (1989).

Rooth, D. 2007. Implicit discrimination in hiring: Real world evidence (IZA Discussion Paper No. 2764). Bonn, Germany: Forschungsinstitut zur Zukunft der Arbeit (Institute for the Study of Labor).

Saul, J. Forthcoming. Implicit Bias, Stereotype Threat and Women in Philosophy. In *Women in Philosophy: What Needs to Change?*, ed. by F. Jenkins and K. Hutchison. Oxford: Oxford University Press. (Formerly titled 'Unconscious Influences and Women in Philosophy'.)

Scheman, N. 2002. Though This be Method, Yet the is Madness in it: Paranoia and Liberal Epistemology. In *A Mind of One's Own: Feminist Essays on Reason and Objectivity*, ed. by L. Antony and C. Witt, Cambridge, MA: Westview.

Steinpreis, R., Anders, K., and Ritzke, D. 1999. The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates: A National Empirical Study. *Sex Roles*, 41: 7/8, 509–528.

Siegel, S. 2013. Can Selection Effects on Experience Influence its Rational Role? *Oxford Studies in Epistemology* Vol. 4: 240-270.

Stewart, B. D., and Payne, B. K. 2008. Bringing Automatic Stereotyping under Control: Implementation Intentions as Efficient Means of Thought Control. *Personality and Social Psychology Bulletin*, 34, 1332-1345.

Unkelbach, C., Forgas, J. and Denson, T. 2008. The Turban Effect: The Influence of Muslim Headgear and Induced Affect on Aggressive Responses in the Shooter Bias Paradigm. *Journal of Experimental Social Psychology* 44: 5, 1409-1413.

# Reliable Misrepresentation and Teleosemantics

**Marc Artiga**
Universitat de Girona - LOGOS

BIBLID [0873-626X (2013) 37; pp. 265-281]

**Abstract**
Mendelovici (forthcoming) has recently argued that (1) tracking theories of mental representation (including teleosemantics) are incompatible with the possibility of reliable misrepresentation and that (2) this is an important difficulty for them. Furthermore, she argues that this problem commits teleosemantics to an unjustified a priori rejection of color eliminativism. In this paper I argue that (1) teleosemantics can accommodate most cases of reliable misrepresentation, (2) those cases the theory fails to account for are not objectionable and (3) teleosemantics is not committed to any problematic view on the color realism-antirealism debate.

**Keywords**
Teleosemantics, Misrepresentation, Mental Content, Color Realism, Mendelovici.

In a recent paper, Mendelovici (forthcoming)[1] has argued that it is possible for representational systems to reliably misrepresent certain properties and that tracking theories such as teleosemantics are incompatible with this possibility. That incompatibility is supposed to highlight a significant drawback of tracking theories. In this paper I would like to argue that one of the most popular tracking theories, teleosemantics, is immune to these criticisms.

The paper is organized in the following way. In the first part, I outline teleosemantics and the problem of reliable misrepresentation. Then, I argue that teleosemantics is compatible with most cases of reliable misrepresentation and that those cases teleoseman-

[1] Since at the moment of writing these lines Mendelovici's paper is only available online, I refer to the pages of the electronic version.

Received: 16/01/2013 Accepted: 04/04/2013

tics rules out are not objectionable. In the final section I address the particular example of reliable misrepresentation Mendelovici has in mind, color anti-realism, and discuss whether teleosemantics is indeed committed to color realism, as she suggests.

## 1 Teleosemantics and reliable misrepresentation

Teleosemantics is a naturalistic theory of content.[2] It is based on two key notions: *function* and *sender-receiver structure*.

First of all, the concept of function employed is the so called *etiological* notion, according to which functions are selected effects (Neander 1995). According to this approach, functions are effects of traits that played an important causal role in the process of selection of that trait. For instance, the function of kidneys is to filter wastes from blood, because kidneys were selected for filtering wastes. That was the effect that explains why organisms having kidneys were favored by natural selection.

Secondly, the sender-receiver structure is an abstract model employed in communication theory and signal detection theory. In short, a sender-receiver structure is composed of two systems, a sender (or ‘producer’), which takes some input and produces a state, and a receiver (or ‘consumer’), which takes this state as input and produces an effect.

Teleosemantics puts these two notions together in order to provide a naturalistic account of representation. In a nutshell, the idea is that representations are states that stand between two systems, a sender and a receiver, when they are endowed with certain etiological functions. The function of the sender is to produce a state R (the *representation*) when another state P obtains (the *representatum*). The function of the receiver is to produce an effect (e.g. a behavior) when state R is tokened.

Now, a key question in the debate on naturalistic theories is what determines the content of the representation. What determines the meaning of R? The answer of mainstream teleosemantics is that R

[2] In this paper I heavily rely on Millikan’s version of teleosemantics, which is the most sophisticated and popular view. Nevertheless, most of the claims presented here could be easily accepted by many teleosemantic accounts.

represents the state of affairs P that the consumer system has historically needed in order to perform its function successfully (Millikan 1993; Papineau 1993). In other words: if we look at the evolution of the sender-receiver mechanism, P is the feature that was required for the consumer to act in a fitness-enhancing way. At the very end, the presence of P is what explains that the whole representational system exists (Neander 2012).

Let me illustrate this theory with an example. Male chicken (*Gallus Gallus domesticus*) produce a characteristic call when they find food, which brings other chicken to that location (Evans and Evans 1999). According to teleosemantics, this call is a representation and means something like *there is food around*, since (1) it stands within a sender-receiver structure, composed by the male chicken (sender) and its fellows (receiver) and (2) the presence of food is what causally explains that the interpreting mechanism (the fellow chicks) performed their function successfully (moving to the location where the call is made and ingesting food). In other words: this representational mechanism was useful because signs correlated with the presence of food usually enough. So, the presence of food is what explains the selection of this sender-receiver structure, and hence it is the state of affairs represented by the call.

Now, Mendelovici's objection concerns what she calls 'reliable misrepresentation'. In contrast to standard criticisms of naturalistic theories of content such as Fodor's (1990) misrepresentation (or disjunction) problem, her argument grants that teleosemantics can account for occasional misrepresentation (Mendelovici: 7). The objection concerns systematic or reliable misrepresentation.

In a nutshell, the idea is that a state type R reliably misrepresents P just in case tokens of R represent instantiations of P, most of the time P is not instantiated and this representational relation is somehow robust (i.e. in similar circumstances R would still misrepresent instantiations of P). More precisely, Mendelovici (4) defines reliable misrepresentation as follows:

Reliable Misrepresentation

An organism's representation type R reliably misrepresents some property P iff:

1. Some tokens of R are involved in attributive mental states that represent objects as having property P,
2. Most or all of the relevant objects do not have P,
3. Tokens of R do or would nonveridically represent objects as having P in the same types of circumstances on separate occasions.

Now, Mendelovici's argues that cases of reliable representation are possible but cannot be accommodated with teleosemantics. That is, she claims teleosemantics is incompatible with the possibility of cases of reliable misrepresentation. The reason is quite simple indeed: since teleosemantics claims that the representational content of R is determined by the state that existed in the past and had a causal influence in the selection of the representational system, it seems to be committed to the usual existence of the represented state. Only something that existed could have had this causal influence. But if the represented state has usually existed in the past, then the representation has been true most of the time. So there seems to be some tension between the teleosemantic view and the existence of systematic misrepresentation.

Mendelovici goes on to argue that this incompatibility with cases of reliable misrepresentation has important consequences for certain philosophical discussions, such as the color realism/anti-realism debate. If teleosemantics is incompatible with cases of reliable misrepresentation, then teleosemanticists cannot accept that our color representations are false most of the time. Consequently, Mendelovici concludes teleosemantics is committed to color realism. And that seems to be a problematic consequence because a general theory of representation such as teleosemantics should not be committed to any particular view on the realism/anti-realism debate. So the theory is in trouble.

I think this objection is worth exploring, not only because this reasoning is intuitively compelling, but also because naturalistic accounts have traditionally had a hard time accounting for the possibility of misrepresentation. In what follows, however, I would like to argue that teleosemantics is fully compatible with the possibility of most cases of reliable misrepresentation, and that those cases that cannot be accounted for are not problematic. In the final section, I will consider the relation between teleosemantics and the debate on color realism.

## 2 Accounting for reliable misrepresentation

The first part of Mendelovici's argument tries to settle that teleosemantics is incompatible with a set of possible cases that involve reliable misrepresentation. However, I think that, as stated, Reliable Misrepresentation clearly fails to pin down this set of situations. Indeed, some of the cases that abide by Reliable Misrepresentation are extremely common, and teleosemanticists usually recognize them (Millikan 1984, 1993; Neander 1995).

Cockroaches and crickets, for instance, possess a set of short appendages at the rear of their abdomen called 'cerci'. Cerci are slender filiform hairs sensitive to air movements. When air moves at a high speed and reaches a certain threshold, a set of neurons are automatically activated and cockroaches perform a range of evasive behaviors (Comer and Leung 2004). According to teleosemantics, this neuronal activation means something like *there is a predator around*, since it is the fact that in some cases a predator was around that explains that the utility of this evasive behavior, and hence accounts for the selection of the mechanism. When there was no predator, this behavior was just a waste of time and energy. But, crucially, notice that probably most of the time cerci produce a neuronal activation and hence an evasive behavior when there is no predator around. That means that the cockroach's neuronal states are probably false most of the time. That is not a problem for teleosemantics because it is still true that what explains the selection of the mechanism is the fact that there was a predator around *usually enough*. Natural selection only requires that in certain cases a trait provides a significant advantage; it does not require that this situation be the most common one.

This is not an isolated example. Whenever a false negative is much more significant than a false positive, signals will tend to be produced in many situations when in fact there is no threat (Godfrey-Smith 1996; Skyrms 2010). When a single false negative could be the last one, organisms tend to produce many false positives (Millikan 1993).

Crucially, notice that this representational system seems to fulfill all conditions set up in Reliable Misrepresentation. The representation R (in this case, certain neuronal activity) represents there being danger around (condition 1), in most cases there is no danger (condition 2) and in many similar circumstances R still misrepresents the presence of danger (condition 3). So cases that satisfy Reliable Misrepresentation are perfectly compatible with teleosemantics.

In order for Mendelovici's argument to have some bite, she needs to exclude this sort of cases, which teleosemantics trivially accounts for. In particular, condition 2 should be restricted to representations of objects that *never* instantiate a certain property P. That is, clause 2 must be read as stating that 'all of the relevant objects do not have P'. Indeed, that version fits much better with the particular counterexamples Mendelovici brings forward in her paper. Let us call this modified definition 'Strong reliable misrepresentation'.

Is teleosemantics incompatible with cases that fulfil Strong reliable misrepresentation? Probably not. We can also describe some scenarios in which teleosemantics can satisfactorily accommodate Strong reliable misrepresentation.

Think first about organisms which are endowed with a mechanism for representing the size of objects. This mechanism will produce a different representation depending on the input it receives. Now, while it is extremely plausible that most of the time we represent the right size of objects, there is a vast literature in psychology describing cases in which this mechanism systematically yields inaccurate representations. For instance, in the best-known version of the Ebbinghaus illusion, two circles of identical size are placed near to each other and one is surrounded by large circles while the other is surrounded by small circles; the first central circle then appears smaller than the second central circle.[3] In this case, the selection and existence of a mechanism producing representations of size is explained by the fact that most of the time it produces the right rep-

[3] Apparently, only the ventral pathway is mislead by this illusion; the dorsolateral pathway represents the size of the objects accurately (Jacob and Jeannerod 2003). So, a better description of the case would say that there is a mechanism in the ventral pathway that reliably misrepresents size.

resentations. Nevertheless, there is a representation type R (the state that misrepresents the size of inner circle in the Ebbinghaus scenario) that reliably and systematically misrepresents a certain configuration. This state R, which reliably misrepresents an inexistent size of certain circles, is a by-product of the representational system that has earned its keep in evolution. This is a sort of case involving strong reliable misrepresentation that can be perfectly accommodated within teleosemantics.

There is a second way teleosemantics can account for cases of strong reliable misrepresentation that Mendelovici does consider (although she utterly rejects this option for reasons that will become clear below). A given mechanism could produce P-involving representations because the organism was confronted with instances of P in the evolutionary past and, nevertheless, at a certain time t, P might not be instantiated any more. Toads (*Bufo Bufo*), for instance, dart on any elongated object moving at a certain velocity in the direction of its axis (Ewert 2004). If a toad is grown up in a laboratory, where all the moving black things it sees are nutritious pellets, this toad will be consistently and reliably misrepresenting all its life. We could even make the case more extreme by supposing that flies go extinct, so that flyhood is never instantiated again. In this case, all toads will be reliably misrepresenting flies. Also in this case, teleosemantics is fully compatible with cases of strong reliable misrepresentation.

Since neither reliable misrepresentation nor strong reliable misrepresentation can do the trick, Mendelovici needs a different kind of example. Indeed, she suggests an additional condition that should also be taken into account: the strong reliable misrepresentation must also be *stable*. She suggests that a reliable representation is stable when it lasts for a significant period of time. Accordingly, the proposal is that the sort of examples that might threaten teleosemantics concern cases of strong and stable reliable misrepresentation. She argues that the example of toads (*Bufo Bufo*) does not exemplify a case of strong and stable reliable misrepresentation, because this sort of misrepresentation is unstable. Since in the laboratory nutritious pellets (and not flies) help toads to survive, teleosemantics entails that toads will come to represent soon the presence of nutritious pellets, rather than the presence of flies.

Now, the notion of stability appealed to here should be qualified in an important way: it is not obvious that teleosemantics cannot accommodate cases of strong and stable reliable misrepresentation, as far as this significant period of time is insufficient for evolution to take place. For instance, teleosemantics is compatible with an individual misrepresenting all its life (e.g. the toad in the laboratory discussed earlier). Unless a process of selection occurs, teleosemantics can accommodate cases of strong reliable misrepresentation lasting for a significant period of time. And we know that within a given population, many years or even centuries might pass by before evolution takes place. So, merely adding stability to strong reliable misrepresentation falls short of specifying a counterexample for teleosemantics: the relevant case Mendelovici is after should include a period of time long enough for evolution to take pace.

Now, prima facie it is hard to think of any example that can satisfy these conditions of stability, reliability and systematic misrepresentation. Mendelovici (10) helps us by describing the particular counterexample she has in mind. She claims that teleosemantics is incompatible with the following situation: a representation type R represents a property P, P correlates with Q and Q (but not P) explains why during a significant period of time the organism has acted successfully and why it was selected for. According to her, that would be a case in which R would be reliably misrepresenting Q as a P. Mendelovici argues this possible situation is ruled out a priori by teleosemantics. On teleosemantics, if instantiations of property Q explain the selection of the mechanism that produces R, then R will be representing Q rather than P. So teleosemantics entails that it cannot happen that R represents P and Q explains why R-representations have been selected for.

Mendelovici's objection, then, is that teleosemantics fails to account for cases of strong and stable reliable misrepresentation involving a process of evolution and in which the property that accounts for the selection of the mechanism (Q) is different from the represented property (P). At this point, I completely agree with Mendelovici. This, I think, is the only case teleosemantics cannot account for. The interesting question, however, is whether this example raises any difficulty for teleosemantics. Is this is an objectionable consequence of the theory?

The answer, I think, is clearly negative. First, remember I showed that Mendelovici can only appeal to cases of strong and stable reliable misrepresentation involving a process of selection. Consequently, her objection is that teleosemantics is incompatible with the following case: Q is the property that causally explains the evolutionary success of having a representation R, but R does not represent Q but P.

I agree this possibility is certainly ruled out by teleosemantics; according to the theory, R represents Q iff Q causally explains the selection of the mechanism producing R. However, that does not seem to be an unwelcome result, but just a different way of stating the theory. If R represents whatever feature explains its selection, it cannot happen that a feature explains its selection and it is not represented by R. Why should that be a problem?

The same point can be made in a different way. Every theory, teleosemantics a fortiori, is such that whatever meets the sufficient conditions for being an F according to a theory is an F according to the theory. This is just what sufficient conditions are. In teleosemantics, those sufficient conditions involve a process of reliability and stability for a period sufficient for selection of the sender-receiver configuration. Consequently, it is certainly true that teleosemantics rules out a case in which Q is the property that accounts for the selection of R and R does not represent Q, but any theory giving sufficient conditions for being an F is incompatible with the presence of sufficient conditions without there being an F. Therefore, the fact that there is a case of reliable misrepresentation that is ruled out by teleosemantics should not be regarded as a problematic consequence of the theory. It is just a different way of formulating the key tenets of this approach.[4]

[4] This response is also available to any other version of teleosemantics. For instance, according to Neander's (1995) approach, content is utterly determined by the properties that the organism is able to discriminate. That is, on her view, toads represent something like *there is a black moving thing*. Now, certainly her theory excludes cases in which an organism represents M and it cannot discriminate M. That is an obvious consequence of the theory, but it is not clear why she should be worried about it.

## 3 Metaphysical commitments of teleosemantics

I just argued that cases of reliable misrepresentation do not threaten teleosemantics. Nonetheless, I think there is a deep and intriguing issue underlying Mendelovici's reasoning that should be addressed by teleosemantics. The interesting question she is trying to raise (which, as I will argue below, she wrongly expressed in terms of reliable misrepresentation) is whether teleosemantics is compatible with certain anti-realist positions, e.g. color eliminativism. This is one of the problems pointed out by Mendelovici's paper I would like to turn to in the remainder.[5]

First of all, notice that, strictly speaking, teleosemantics as such is compatible with the denial of realism about any entity. For instance, teleosemantics can be true at a possible world *w*, even if no organism exists at *w*, or even if evolution has never taken place at *w*. What the theory claims is that there are representations at a world *w if and only if* certain processes occur (involving some systems, natural selection and so on). So teleosemantics as such has no realist implications.

Mendelovici, however, tries to put some pressure on teleosemantic theories in that direction. She argues that if we accept teleosemantics *and* certain empirical claims, the theory has inadequate consequences concerning the debate between realism and anti-realism. In particular, she claims that accepting teleosemantics 'would force us to be realists about properties represented in nonsemantically successful conditions, where realism about property P is the view that P is instantiated' (Mendelovici: 18). Now, does teleosemantics force us to be realist about properties represented in nonsemantically successful conditions? I think the answer is affirmative, but again I doubt granting this point results in any problematic consequence for

[5] Mendelovici (16-7) puts forward another argument, which I think can be easily defeated given the results of the previous discussion. First she distinguishes veridicality from reliability: a mechanism is veridical if it yields the right result and it is reliable if it tends to produce the same result, regardless of whether it is veridical. Her 'psychological argument' claims that if a theory is unable to account for reliable misrepresentation, it cannot maintain the useful distinction between veridicality and reliability. Now, since I have already shown that teleosemantics can accommodate many cases of reliable misrepresentation, it should be obvious that it can also make this distinction.

the theory. Let me explain.

In teleosemantics 'P is represented in non-semantically successful conditions' should be spell out as the claim that P was the property that accounted for the selection of the sender-receiver system. Hence, the problem should be cashed out as follows: if teleosemantics is right and P is the property that accounts for the existence and selection for the system, then one is committed to the (past) existence of P.[6] Again, this conditional seems true, but also entirely plausible. If a property P accounted for the existence of the representational system, P must have been instantiated somewhere.

Indeed, this inference is not only plausible, but it describes a standard way of reasoning in science. A clear example can be found in research on arachnophobia (the fear of spiders and other arachnids). Some studies suggest that evolution might have equipped mammals with a strong predisposition to react fearfully to spiders (Öhman and Mineka 2001). One of the standard criticisms to this proposal is that only 0.1 percent of the 35.000 different kinds of spiders in the world are poisonous, so probably having a predisposition to react fearfully to spiders did not constitute any significant selective advantage for mammals (Gerdes et al. 2009). The debate, then, assumes that if humans possess a mechanism for producing fearful reactions to spiders, then a sufficient number of poisonous spiders must have existed in the past. Hence, scientists accept that if we are endowed with a mechanism for representing or behaving towards P, then P must have existed in the past. That looks like an impeccable scientific reasoning.

Similar arguments are also common in philosophy, e.g. in relation to radical conceptual nativism (Fodor 1975, 1998). Some people have suggested that human concepts like CARBURETOR or TELEVISION cannot be innate because if they were, we would have to accept that there were carburetors and televisions at the time our ancestors evolved (Sterelny 1989; Prinz 2002: 229). Again, it is assumed here that the truth of some claims about conceptual content

[6] It is worth pointing out that if teleosemantics is committed to realism about an entity, it is realism about an entity in the evolutionary past. As we saw, teleosemantics is compatible with the possibility of P-involving representations in cases where P has gone extinct.

commits one to certain ontological claims.

Therefore, one should not try to devise a general argument against drawing metaphysical conclusions from theories of meaning and certain empirical claims about the current representational capacities of organisms. Many discussions in science and philosophy take for granted that if we have good reasons for thinking that an organism has build-in a mechanism for representing a set of properties P, this set of properties P existed in the past.

In order to make her objection to teleosemantics more compelling, Mendelovici focuses on a particular case, in which this general way of arguing seems to go astray: color. More precisely, she argues that if we accept teleosemantics, the following inference could be carried out:

P1 I have experiences of redness.
P2 My experiences of redness at least sometimes occur in (nonsemantically) successful conditions.
P3 If I have experiences of redness in (nonsemantically) successful conditions, then realism about redness is true.

———————————————

C Realism about redness is true.

She claims that realism about a certain entity like colors should not be so easy to get, so given the strong intuitive support for P1 and P2, P3 (which directly derives from teleosemantics) should be abandoned. The objection, then, is that teleosemantics warrants certain inferences and conclusions in the realist debate that a theory of content should not allow.

Before directly addressing this particular version of the objection, let me clarify the first premise, which I think is unduly imprecise. P1 could be interpreted (at least) in three different ways: as a claim about representational content (I have experiences about redness), about phenomenal properties (I have experiences instantiating phenomenal redness) or about the two (I have an experience about redness instantiating phenomenal redness). Now, remember that this argument is supposed to make explicit a consequence of endorsing teleosemantics and this naturalistic approach is exclusively concerned with representational content (see Millikan 1984, Neander 2012). Teleosemantics as such is a theory of content, and it is silent concerning the relationship between representational content and phenomenal properties. Therefore, in that particular case, P1 should be read as stating that experiences are about redness. That is, P1 is a claim about representational content.

Indeed, notice if P1 were interpreted as stating that experiences instantiate a certain phenomenal property (redness), P1-P3 would not entail C without additional and controversial assumptions about the connection between phenomenal and representational properties, which are clearly not made by teleosemanticists and would require independent support.

Having clarified premise P1, let me argue why I think Mendelovici is right in her formulation of the inference, but (again) teleosemanticists can happily accept this result.

First of all, we saw that drawing certain ontological claims from a theory of meaning plus certain empirical claims is generally regarded as valid. So what is wrong with this reasoning? Some people will probably find this particular inference objectionable because of the a priori status of P1 and P2. Since P1 and P2 seem to be priori and C is clearly a posteriori, if we accept that P1-P3 entail C, we will be entitled to conclude a substantive and a posteriori claim about the world (color realism) from certain a priori claims and teleosemantics. I think this is precisely what worries Mendelovici, since she thinks that P1 can be established by introspection alone, P2 is uncontroversial and she accuses teleosemantics of enabling one to become a realist about a certain entity 'without any empirical examination of objects' (Mendelovici: 17).[7]

[7] Here is another quote where Mendelovici makes clear that her main objection concerns the fact that premises P1 and P2 are a priori while the conclusion is a posteriori: 'But if tracking theories are correct, then in order to establish realism about a represented property P, we needn't check the world for evidence of instances of P. We can instead check ourselves for nonsemantically successful instances of the representation of P. (Mendelovici, forthcoming, 18; so also footnote 19). Nonetheless, let me point out that, at the same time, Mendelovici claims that she is not concerned about the a priori status of the premises (20). Again, if this is true, I fail to see why this inference is problematic.

However, this is surely not warranted by teleosemantics. Teleosemantics is an externalist theory about content, so P1 and P2 are a posteriori claims through and through. What kind of property I am representing with a red experience and what kind of situations are nonsemantically successful conditions (i.e. what sort of situations accounted for the selection of the mechanism) are hard empirical questions that should be resolved by science. Consequently, even if teleosemantics is right, a considerable amount of empirical knowledge must be gathered before anything like C can be established. Certainly, if we accept that color-experiences are representations (as they probably are) then, teleosemantics is committed to there being a property they are supposed to represent. However, what kind of entity we are committed to is something that should be discovered by empirical research.

Of course, if one assumes that the content of mental states can be discovered through introspection alone (as Mendelovici seems to suggest in 18), then this claim is in tension with teleosemantics. In general, externalist theories threaten the privileged access we seem to have to the content of our own mental states. But we already knew that externalist theories (and teleosemantics among them) are in tension with certain internalist intuitions, so on this interpretation there is nothing new about Mendelovici's argument (see Boghossian 1997). Furthermore, notice that if this is what the argument intends to show, there is no specific objection to teleosemantics or tracking theories: any externalist theory of content has this difficulty. Accordingly, a defense will have to come from externalism, rather than from teleosemantics.

Finally, let me conclude by directly addressing Mendelovici's main question: if we assume teleosemantics and grant everything I accepted in this paper (including the inference from P1-P3 to C), is teleosemantics still compatible with color eliminativism? It clearly is. If science discovers that there is nothing our color experiences have been tracking, then teleosemantics has to say that the mechanism that produces our color experiences is not a representational mechanism. That is, it is possible that color experiences are not representational states. There are many alternative evolutionary explanations for the existence of the mechanism: evolutionary drift, sprandels, and so on. Of course, I am not saying that this is an attractive view;

but the objection was based on the incompatibility of teleosemantics and color eliminativism, not on the plausibility of the latter. What we had to show is that accepting teleosemantics and granting the validity of the inference described earlier does not make teleosemantics incompatible with color eliminativism.[8]

Finally, let me make a general comment on Mendelovici's dialectical strategy. In this paper, I have addressed the question of reliable misrepresentation and the question of color realism separately. But Mendelovici's strategy is to use the first debate in order to conclude something about the second. In particular, she argues that (1) teleosemantics is incompatible with color-experiences reliably misrepresenting and then that (2) if color eliminativism is true, then color-experiences reliably misrepresent. If (1) and (2) were true, then teleosemantics would be incompatible with color eliminativism. However, in the first part of the paper I showed that (1) is false and I just presented an argument suggesting that (2) is false as well. According to teleosemantics, if there are no colors (if there is nothing our color experiences have been tracking), then the mechanism producing color experiences is not a representational system and, consequently, color experiences are not representations. And, of course, if experiences do not represent anything, they cannot misrepresent either. So, Mendelovici is wrong in assuming that the only way for teleosemantics to be compatible with color eliminativism is by accommodating reliable misrepresentation. Consequently, the question about reliable misrepresentation and the question about color eliminativism should be clearly distinguished. This is the reason I have addressed (1) and (2) separately.

## 4 Conclusion

In conclusion, I argued that teleosemantics can account for most cases of reliable misrepresentation and that those particular instances

[8] Of course, I am not denying that certain views are incompatible with Teleosemantics. For instance, probably one cannot coherently hold at the same time that (1) teleosemantics is true, that (2) color experiences are representational, and that (3) there is nothing color experiences have been tracking in the evolutionary past. However, prima facie this sort of incompatibilities do not seem to be a problem. Thanks to a referee for pressing me on this issue.

that cannot be accommodated do not pose any significant problem for the theory. On the other hand, Mendelovici's general objection against drawing ontological conclusions from a theory of content (plus certain empirical claims) seems to be wrong-headed. In sum, teleosemantics is neither threatened by cases of reliable misrepresentation, nor by any metaphysical consequence of the theory.[9]

Marc Artiga
Departament de Filosofia
Facultat de Lletres
Universitat de Girona
Plaça Ferrater i Mora, 3
17071 Girona
marc.artiga@gmail.com

## *References*

Boghossian, Paul. 1997. What an externalist can know a priori. *Proceedings of the Aristotelian Society* 97 (2): 161-175.

Comer, Christopher and Leung, Vicky. 2004. The Vigilance of the Hunted: Mechanosensory-Visual Integration in Insect Prey. In *Complex Worlds from Simpler Nervous Systems*, ed. by F. R. Prete. Cambridge: MIT Press.

Evans, Christopher and Evans, Linda. 1999. Chicken food calls are functionally referential. *Animal Behavior* 58(2):307-319.

Ewert, Jörg-Peter. 2004. Motion perception shapes the visual world of amphibians. In *Complex Worlds from Simpler Nervous Systems*, ed. by F. R. Prete. Cambridge: MIT Press.

Fodor, Jerry. 1975. *The Language of Thought, Cambridge*, Cambridge: Harvard University Press.

Fodor, Jerry. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

Fodor, Jerry. 1990. *A Theory of Content And Other Essays*. Cambridge: MIT Press.

Gerdes, A., Uhla, G., Alpersa, G. 2009. Spiders are special: fear and disgust evoked by pictures of arthropods. *Evolution and Human Behavior*, 30: 66–73.

Godfrey-Smith, Peter. 1996. *Complexity and the Function of Mind in Nature.* Cambridge: Cambridge University Press.

Jacob, Pierre and Jeannerod, Marc. 2003. *Ways of Seeing: The Scope and Limits of Visual Cognition.* Oxford: Oxford University Press.

Mendelovici, Angela. Forthcoming. Reliable misrepresentation and tracking the-

[9] I would like to thank Manolo Martínez, Miguel Àngel Sebastian, David Pineda and an anonymous referee for helpful comments and incisive criticisms. This work was supported by the scholarship BES-2008-005255 from the Spanish Ministry of Economy and Competitivity (MEC), the Research Projects 'The Naturalization of Subjectivity' (ref. FFI2010-15717), 'Modal Aspects of Materialist Realism' (ref. HUM2007-61108) and Consolider-Ingenio project CSD2009-00056.

ories of mental representation, *Philosophical Studies.*

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

Millikan, Ruth. 1993. *White Queen Psychology and Other Essays for Alice*, Cambridge: MIT Press.

Neander, Karen. 1995. Malfunctioning and Misrepresenting. *Philosophical Studies* 79: 109-141.

Neander, Karen. 2012. Teleological Theories of Mental Content. *Stanford Encyclopedia of Philosophy.*

Öhman Arne and Mineka, Susan. 2001. Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review* 108(3): 483-522.

Papineau, David. 1993. *Philosophical Naturalism.* Cambridge: Basil Blackwell.

Prinz, Jesse. 2002. *Furnishing the Mind: Concepts and their perceptual basis*. Cambridge: MIT Press.

Skyrms, Bryan. 2010. *Signals, Evolution and Learning.* Oxford: Oxford University Press.

Sterelny, Kim. 1989. Fodor's Nativism, *Philosophical Studies* 55 (2): 119-141.

# A Puzzle about Disagreement

**Víctor M. Verdejo**
University of Santiago de Compostela

BIBLID [0873-626X (2013) 37; pp. 283-297]

**Abstract**

A well-known (roughly Fregean) strategy for dealing with Kripke's 1979 Pierre-puzzle is to appeal to differing senses or modes of presentation in the characterization of Pierre's beliefs. However, differing senses or modes of presentation in the characterization of an agent's beliefs conceal, in this context, another equally challenging puzzle about disagreement. Apparently therefore, theorists are required to pay attention to both sorts of puzzles in order to offer a satisfactory solution to the Pierre case.

**Keywords**

Disagreement; Kripke's puzzle; mode of presentation; rationality

## 1 A celebrated puzzle about belief

Kripke 1979 introduced a celebrated puzzle about belief. In one of its versions (1979: 254-59), the puzzle concerns Pierre's assent to 'Londres est jolie' and dissent to 'London is pretty'. Pierre does not realize that 'Londres' and 'London' are names of the same city. Following highly plausible disquotational and translation principles[1], in the envisaged scenario, Pierre, a perfectly rational subject, believes and disbelieves that London is pretty. However, rational subjects do not let contradictory beliefs pass. We can express this puzzle as the paradoxical truth of both (1) and (2), which state a two-place belief (B) relation between Pierre and two contradictory propositions containing the monadic predicate 'P' for 'being pretty' and the constant '*l*' for the city:

[1] The disquotational principle serves to derive a subject's beliefs from a subject's linguistic assent and the translation principle is used for the French-to-Eng

*Disputatio*, Vol. V, No. 37, November 2013

Received: 16/02/2013 Revised: 05/04/2013 Accepted: 03/07/2013

(1) $B(\text{Pierre}, P\mathit{l})$
(2) $B(\text{Pierre}, \neg P\mathit{l})$

Scholars have proposed different solutions to Kripke's original considerations. In this paper, I undertake the task of showing that one kind of preferred (roughly) Fregean strategy to solve the puzzle fails. The general strategy consists of analyzing Pierre's attributed beliefs as involving different senses, modes of presentation or analogous intensional categories, for the expressions 'Londres' and 'London', respectively. Many authors, who differ in much else, have shown confidence about the fruitfulness of following such strategy in order to explain away the puzzle.[2] Differing senses or modes of presentation are taken to save Pierre's rationality. On this reading, there is no contradiction in Pierre's beliefs but simply different, perfectly compatible, beliefs. We may state this widely trusted strategy in terms of (3) and (4), where different context-sensitive senses or modes of presentation '$l_1$' and '$l_2$' are assigned to Pierre's belief-contents in France and England, respectively:

(3) $B(\text{Pierre}, Pl_1)$
(4) $B(\text{Pierre}, \neg Pl_2)$[3]

(3) and (4) may be taken to involve senses or modes of presentation for 'Londres' and 'London' in ways that contrast with the descriptivist standpoint that pervades Kripke's own consideration of senses and

lish translation of belief reports (see Kripke 1979: 248-254). I will not discuss these principles here (but see Section 3).

[2] Variations of this line of reasoning can be found in Chalmers (2011: 611-12), Crimmins (1992: 161-63), Richard (1990: 179-83), Salmon (1986: 129-32), Schiffer (1992: 507-10), Sosa 1996 or Zalta (1988: 189-96), among many others.

[3] For expository purposes, I am here passing over analyses of belief reports in terms of triadic relations between a subject, a Russellian proposition, and a mode of presentation. Little of the argument in the main text bears on this variation of the (roughly) Fregean strategy under consideration. In particular, the argument here presented applies equally to those (more Russellian) views that would deny that the modes of presentation in belief reports are a feature of the propositional contents in that-clauses (e.g., Crimmins 1992, Salmon 1986, Schiffer 1992). Once this is clear, I will drop the parenthetical caveat 'roughly' in what follows.

modes of presentation. Nonetheless, this strategy, however exactly understood, fails to come to grips with a standard notion of disagreement. In so failing, I will argue, it apparently changes Kripke's puzzle about belief for another puzzle about disagreement.

## 2 A puzzle about disagreement?

Disagreement is approached from a variety of angles in the recent literature. According to a baseline and largely undisputed characterization, however, disagreement is understood in terms of the acceptance or belief of contradictory propositions. More precisely, it is taken as a necessary (although very probably not a sufficient) condition for $S_1$ and $S_2$ to disagree about whether *p*, that $S_1$ and $S_2$ hold or accept contradictory *p*-beliefs or belief-contents. From this, we can make explicit the principle of disagreement PD:

> (PD) Two subjects $S_1$ and $S_2$ disagree only if, for any *p*, $B(S_1, p)$ and $B(S_2, \neg p)$[4]

In the light of PD, (3) and (4) can be shown to be implausible. The reason is that they save the rationality of Pierre at the price of jeopardizing the possibility of general disagreement between people in France that assent to 'Londres est jolie' and people in England that assent to 'London is not pretty'. According to (3) and (4) and PD, two subjects being in an entirely analogous scenario as that of Pierre's (in France and England, respectively) would not count as genuinely disagreeing. They will count simply as believing different, not really contradictory, propositions.

To make the point fully explicit, let us call the relevant subjects Antoine and Anthony. Let us imagine that Antoine is exactly like Pierre before moving to England. Antoine is a normal French speaker who hears beautiful things of a distant city called 'Londres'. On the other hand, let us suppose further that Anthony is a fluent speaker of English who lives in exactly the same unattractive quarter as Pierre's in London. Anthony is just like his Londoner neighbours

[4] PD is both intuitively plausible and an explicit theoretical principle of the reflection on disagreement across a number of philosophical issues. More on PD in Section 3.

regarding knowledge and linguistic capacities. As a result, Antoine has an inclination to assent to 'Londres est jolie' whereas Anthony has an inclination to assent to 'London is not pretty'. We may think of Antoine and Anthony's assents as being as enduring and definitive as in the case of Pierre. Now, according to our intuitive understanding of disagreement, Antoine and Anthony should be taken to disagree with each other about the beauty of London. This is a hardly questionable diagnosis. It should be a matter of course that Antoine and Anthony do disagree with each other about London if anyone ever did. Since PD expresses a necessary condition on disagreement, and using the same notation as in (1) and (2), Antoine and Anthony should be described in terms analogous to (5) and (6):

(5) B(Antoine, P*l*)
(6) B(Anthony, ¬P*l*)

If this is correct, a puzzle about disagreement immediately ensues for the Pierre case. Defenders of analysis of accounts along the lines of (3) and (4) seem far from being capable of accommodating situations such as the ones expressed by means of (5) and (6). In particular, if Pierre differs in his French and English thoughts about London, then, for the same reasons, Antoine and Anthony, who are exactly like Pierre in the relevant respects, should be taken to differ regarding their beliefs about the city. The puzzling situation can be stated by means of a dilemma: either Pierre is rational in believing what (3) and (4) specify and we therefore fail to account for the (potential) disagreement between speakers of different languages such as Antoine and Anthony; or we account for this disagreement at the price of giving up Pierre's rationality.

## 3 The assumptions of the puzzle about disagreement

It is worth considering explicitly all the assumptions involved in the new version of Kripke's puzzle so far. The puzzle about disagreement originates from endorsement of the following:

i) There is a puzzle about the rationality of Pierre's beliefs which has to be explained or explained away.

ii) There are (at least basic) cases of intercultural disagreement among people speaking different languages (like the one exemplified by Antoine and Anthony).

iii) PD is true.[5]

Now, the foregoing considerations can be summarized in terms of a conditional claim as follows: if i) through iii) are true, then any Fregean solution to the rationality of Pierre's beliefs–along the lines of (3) and (4) — leads us to the impossibility of accounting for (basic) cases of disagreement — along the lines of (5) and (6). If true, I believe that this conditional claim is likely to bring with it significant consequences for the analysis of rationality puzzles and the debate between Fregeans and Millians regarding the meaning of proper names. Here, however, I wish to be neutral about what these consequences exactly are. In this paper, I am only concerned with showing that the conditional claim just espoused is in fact true. In this section, I would like to introduce some clarifications regarding the assumptions i)-iii). Since they are assumptions, I will not try to argue for them in what follows but only to motivate them and to show their initial plausibility and relevance for the present discussion.

First, i) need not be universally accepted. For instance, the truth of the disquotational and translation principles which, according to Kripke, give rise to the puzzle might be questioned. Kripke himself safeguarded his 1979 argument against objections to the translation principle by introducing the Paderewski case in which only one language is involved (Kripke 1979: 265-266). If the disquotational and translation principles are not true, then arguably the puzzle would not arise in the first place. Note however that it is not less true that the puzzle can be made to arise independently of these principles (e.g. Salmon 1986: 130, Sosa 1996: 384-5). More importantly for present purposes, whether or not we accept the disquotational and translation principles or other principles that originate the puzzle, all the proponents of the Fregean solution to the puzzle here under scrutiny would very clearly and resolutely accept i). One may therefore

[5] As we will see in due course (Section 4), a further assumption iv) will be added to the list. We can rest content with the analyses of these three assumptions for now.

just assume that Pierre's puzzle about rationality is possible independently of any principles.

One might, on the other hand, be tempted to defend that Kripke's puzzle is not exactly a puzzle about rational belief but more clearly a puzzle about our theories of belief ascription. In this regard, I am sympathetic to Kit Fine's suggestion that both the problems arising from the theoretical correctness of (1) and (2) and their truth are intrinsically connected (Fine 2007: 89). For present purposes, nonetheless, it is enough to see that the proponents of the Fregean solution under consideration have also agreed in accepting the need to address the puzzle in the terms just described in Section 1, that is to say, in terms of the paradoxical truth of (1) and (2) for a Pierre who 'would *never* let contradictory beliefs pass' (Kripke 1979: 257, emphasis his).

As regards ii), I do not believe that philosophers would easily find reasons to doubt it, at least initially. If ii) is correct, and disagreement concerns at least basic ordinary cases, then it would follow that the Antoine and Anthony case is a paradigmatic case of disagreement. Their story is not weird or confused in any way. They are subjects that show commitment to and hence are said to believe contradictory propositions. There might be similar, more involved related cases for which there is actually no disagreement. But that is of course no reason to doubt that the Antoine and Anthony case is, in normal settings, a basic case of disagreement. This outcome is not itself troublesome or paradoxical. Antoine and Anthony need not ever interact with each other. For that matter they need not even be coetaneous. However, there does not seem to be any intuitive or philosophical basis that would disallow concluding that they disagree if anyone ever did.[6]

Of course, cases of disagreement might be controversial in a number of *other* ways. For instance, it is controversial whether one can explain all the epistemic puzzles that seemingly reasonable disagreement brings with it (see Feldman 2006). Philosophers' intu-

[6] As in the case of i), ii) might be taken to be the result of some disquotational and translation principles. As before, however, we do not need to embrace this commitment. We may simply assume that Antoine and Anthony disagree independently of any principles, granted that there is nothing especial or wanting about their acquiring contradictory beliefs in the way they do.

itions might run quite disparately for these specific cases. Please note that ii) is not controversial in this way. There does not seem to be, at least initially, disparate intuitions about whether persons in the situation exemplified by Antoine and Anthony would be disagreeing. Acceptance of this case only involves acknowledgement of the existence of (basic cases of) intercultural and multilingual disagreement. This is not to say that ii) is undoubtedly true. Misleading intuitions abound elsewhere. However, it should strike us as a major philosophical finding if it turned out that ii) was, in fact, false or suspicious.

Finally, it is fair to take iii) as one of the foundational assumptions of the contemporary literature on disagreement. Disagreement has recently come to the philosophical scene in a variety of challenging ways. We clearly lack a general theory for such different fields as moral or religious disagreement, disagreement about matters of taste, or the epistemic import of peer disagreement, to name a few. The polemics regarding the notion of disagreement have been located, in many cases, on the examination of which conditions are actually sufficient for disagreement. All the same, PD expresses a largely unquestioned necessary condition about propositional attitude disagreement, or more precisely, about belief disagreement.[7] Authors who differ in much else would have no query in accepting iii). The intuition that underlies PD is that genuine (belief) disagreement requires contradictory or other sorts of incompatible beliefs. Acceptance of PD drives a number of controversies including the one confronting contextualism and relativism in the philosophy of language (e.g. MacFarlane 2007, Richard 2011: 425). There is also a wide consensus about the truth of PD in discussions on the epistemic status of our beliefs (see e.g. Feldman and Warfield 2010 collection). PD is also present in other kinds of analyses, such as the ones that deepen into the prescriptive nature of disagreement (e.g. Ridge 2013). It seems therefore reasonable to take assumption iv) also as well-motivated in the present discussion.

[7] See however Sundell 2011 for a rejection of a linguistic version of PD based on an analysis of linguistic denial.

## 4 Rationality and disagreement

Once the above clarifications are made, full appreciation of the force of the foregoing considerations requires exploring the way in which a theorist may try to maintain (3) and (4) and still possibly accommodate (5) and (6). Authors have suggested a number of specific ways in which this might be done. The key thought would be that the beliefs specified by means of (3) and (4) are suitably related to the beliefs specified by means of (5) and (6), respectively. By appealing to such relation — it may be argued — a sense can be made of Pierre's holding non-contradictory propositions (and therefore being rational) and, at the same time, Antoine and Anthony's holding contradictory propositions (and therefore genuinely disagreeing). The target relation may be understood, for instance, in terms of co-referentiality (e.g. Zalta 1989, Heck 1995), similarity of Fregean thoughts (Forbes 1987), determination of the same representational type (Richard 1990: Chapter 3) or coordination (Fine 2007, Chalmers 2011), to name a few. For simplicity's sake, let us focus on Chalmers's 2011 more recent account in terms of coordination which, unlike others, explicitly addresses disagreement in the context of a Fregean approach to belief ascription.

Chalmers's account is based on enriched propositions, a theoretical construct belonging to two-dimensional semantics. The exact nature of these propositions need not concern us here. For present purposes, it is enough to know that enriched propositions are complex structures constituted by primary intensions (i.e., Fregean thoughts and constituent senses for present purposes) and secondary intensions (i.e., in this context, Russellian or purely referential propositions or propositional-constituents). On this account, therefore, propositional contents (and thus the objects of beliefs) are hybrids made out of pairs of Fregean thoughts and Russellian propositions.

As advanced, Chalmers 2011 characterizes agreement/disagreement in terms of coordination. To understand this, we need to note that Chalmers distinguishes between endorsement and belief (2011: 619-21). On this view, belief is analysed in terms of endorsement plus coordination. Thus, for *p* to be a(n) (enriched) proposition that S believes, S need not endorse *p*. It is enough, Chalmers contends, that S endorses another proposition, *p*', which is suitably coordinated with *p*. This results in a specific thesis about disagreement. In particular, Chalmers suggests that 'two people disagree when one believes a proposition *p* and the other believes ¬*p*, which requires that one endorses a proposition coordinate with *p* and the other endorses a proposition coordinate with ¬*p*' (2011: 619). Note that this characterization is perfectly compatible with our fundamental principle (PD).

Now, Chalmers (2011: 611-12) claims that Kripke's case should therefore be handled by means of the two-place endorsement (E) relation between subjects and propositions and the two-place coordination (C) relation between propositions. Following the here presented notation, we can state this proposal for analysing the Pierre case via (7) and (8):

(7) E(Pierre, $\mathrm{P}l_1$) & C($\mathrm{P}l_1$, $\mathrm{P}l$)
(8) E(Pierre, $\neg\mathrm{P}l_2$) & C($\neg\mathrm{P}l_2$, $\neg\mathrm{P}l$)

(7) says that Pierre endorses the (enriched) proposition that contains the (enriched) intension associated with 'Londres' ($\mathrm{P}l_1$) and (8) says that he endorses the (enriched) proposition that has the (enriched) intension associated with 'London' as a constituent ($\neg\mathrm{P}l_2$). In turn, $\mathrm{P}l_1$ is coordinated with $\mathrm{P}l$; whereas $\neg\mathrm{P}l_2$ is coordinated with $\neg\mathrm{P}l$. Note that, on this account, we are in an apparently adequate position to capture both the intuition that Pierre is rational and the intuition that people in France and England — such as Antoine and Anthony — may disagree by assenting to 'Londres est jolie' and to 'London is not pretty', respectively. On the one hand, we can allegedly save Pierre's rationality because he is viewed as endorsing different, perfectly compatible, propositions. On the other hand, we can apparently account for the disagreement of French and English contenders on the grounds that endorsed propositions are coordinated with suitable propositions that bring about public belief and disbelief.

Remarkably, as (7) and (8) show, Chalmers's proposal proceeds by splitting up issues regarding rationality, which are understood in terms of the propositions people endorse (in the technical sense of 'endorse'), and issues about disagreement, which are understood in terms of the propositions people believe (in the technical sense of 'believe'). Indeed, Chalmers explicitly claims that, unlike believed propositions, endorsed propositions are the ones 'constitutively related to rationality' (Chalmers 2011: 612).

Chalmers's solution, and any solution along the lines of (7) and (8) is, however, clearly unconvincing. In the first place, according to (7) and (8), Pierre is said to believe *and* disbelieve that London is pretty. The solution therefore appeals to a peculiar sort of 'irrational' or contradictory belief. (7) and (8) describe an odd situation for a rational *believer*. And this is regrettable on the face of the fact that Pierre is, by assumption, a perfectly rational logician and philosopher. How could then Pierre believe contradictory propositions?[8] In this context, therefore, Chalmers's proposal would seem, at most, a redescription of and certainly not a successful solution to the puzzle. Beliefs attributed to Pierre in the light of his assent to 'Londres est jolie' and dissent to 'London is pretty' was supposed to give rise to the puzzle. The puzzle was, precisely, that one cannot say whether Pierre believes or not that London is pretty. It was an assumption of the puzzle — indeed, it was the very puzzle — that Pierre should *not* be said to believe *and* disbelieve that London is pretty.

> 'To reiterate, this is the puzzle: Does Pierre, or does he not, believe that London is pretty? It is clear that our normal criteria for the attribution of belief lead, when applied to *this* question, to paradoxes and contradictions.' (Kripke 1979: 259, emphasis his)

The main paradox is, of course, that Pierre, according to such normal criteria, is said to hold contradictory beliefs.

Philosophers might be willing to protest when faced with these considerations. Perhaps the only way in which we can get rid of the puzzle about belief, these authors might argue, is by appealing to a modified or technical notion of belief. Perhaps there is a sense in which our preferred notion of belief might be correctly considered as 'irrational', in the specific sense of permitting contradictory beliefs. If this is correct, we might have reasons to abandon the Kripkean assumption that rational subjects cannot hold contradictory beliefs. It might thus be defended that we are not really forced to determine whether Pierre believes that London is pretty or not.

[8] The result is all the more surprising if we reflect on the fact that Pierre, if acquainted with the facts, would arguably not hold *any* of the beliefs attributed by means of (7) and (8) (Goldstein 2009).

The situation, however, is patently more problematic than the foregoing remarks suggest. It is not only that accounts along the lines of (7) and (8) try to solve Kripke's paradox by resorting to a peculiar contradictory notion of belief. In addition, analyses in terms of (7) and (8) are themselves paradoxical regarding the notion of disagreement.

To show this, let us suppose that it makes sense to modify our notion of belief in such a way that Pierre may let contradictory beliefs pass, after all. Let us accept, following Chalmers, that rational subjects may hold contradictory beliefs, according to the new propounded notion of belief. Now, independently of whether such a notion of belief is plausible or justified, the analysis forces us to conclude what we cannot in any way accept, namely, that Pierre disagrees with himself with respect to *p*. According to (7) and (8), Pierre believes both that *p* and that $\neg p$. By Chalmers's own admission, this is precisely the way in which we may characterize disagreement. It follows that Pierre disagrees with himself about the beauty of London. This has the form of a reductio: rational subjects, such as the brilliant logician and philosopher Pierre, do not disagree with themselves.

Defenders of the analysis here under consideration might think it useful at this point to concentrate on the fact that even if PD — and Chalmers's own characterization — provide an intuitively correct sense of disagreement, something can be said in the case of Pierre to argue that, in this particular case, contradictory beliefs are not sufficient — even if perhaps necessary — for genuine disagreement. Thus, advocates of (7) and (8) might be willing to appeal to further conditions, besides contradiction in beliefs, in order to resist the conclusion that Pierre disagrees with himself.

Once arrived at this point, however, the paradoxical impetus of the Pierre case shows its full force. Defenders of accounts along the lines of (7) and (8) need something they cannot achieve: they need to conclude that Pierre does not disagree with himself and, at the same time, to characterize Pierre's beliefs in such a way that subjects being in exactly the same situation as Pierre does for holding the beliefs he does — such as Antoine and Anthony — would nevertheless be disagreeing. In short, further conditions besides PD would be conditions that either make Pierre disagree with himself, or conditions that do not explain genuine disagreement between French- and English-speakers. It would seem that theorists cannot have it both ways, no matter what amount of reasonable conditions they add to PD.[9]

Authors willing to hold to (7) and (8) may perhaps try to bite the bullet. They may accept that their account for the Antoine and Anthony case forces them to admit that Pierre, a perfectly rational subject, disagrees with himself. However, they may reason, this is only regrettable on the face of however doubtful intuitions. More precisely, the foregoing argument assumes that reflexive disagreement of the sort expressed in (7) and (8) is untenable. It may turn out that this assumption is wrong. Rational subjects may disagree with themselves in the technical sense of disagreeing. Just as in the case of the notion of 'irrational' belief, therefore, the technical notion of disagreement may allow for intuitively odd or unfamiliar but perfectly legitimate results. We may have to admit therefore the possibility of reflexive disagreement after all.

It is true that the argument at this point assumes the impossibility of reflexive or self-disagreement. We could add this assumption to the set introduced in Section 3 in terms of iv):

iv) Rational subjects do not disagree with themselves as to whether *p*, for any given *p*.

An objector may therefore wish to defend rejection of iv) by appealing to the technical sense of disagreement required for dealing with the Pierre puzzle. The puzzle, it may be argued, forces us to refine our intuitive, unexplained notion of disagreement, the one that would seem inextricably tied to iv).

The prospects of this line of reply seem however dim. The alleged technical or explicated sense of disagreement is not only at odds regarding our intuitions about rational belief. That is to say, it not only involves accepting that contradictory beliefs may nonetheless be rational. It is also unsatisfactory from the point of view of a

[9] The presumption is of course that this outcome has nothing particular to do with Chalmers's specific way of addressing the issue. The problem seems to arise likewise for any other candidate analysis that is Fregean in the relevant sense, namely, in the sense of introducing different intensional categories for making rational Pierre's beliefs or attitudes in France and England.

minimally plausible conception of *rational disagreement*. For, unlike perhaps belief, disagreement clearly involves holding oneself responsible for or being normatively committed to the truth or correctness of the contents or attitudes figuring in the disagreement.[10] It should be out of the question that rational subjects, such as our leading logician Pierre, cannot hold contradictory normative commitments of this sort. But this is precisely what we are forced to allow if we hold to analyses along the lines of (7) and (8).

We seem to have therefore powerful reasons to conclude that if one's preferred theory entails rejection of iv), then one's preferred theory should be dismissed as an account of a minimally plausible notion of disagreement. Disagreement, we may say, is not the result of a would-be relation one happens to have towards possibly contradictory beliefs. Disagreement requires subjects who hold themselves responsible for and are normatively constrained by what they believe. This claim is not apodictic. Admittedly, the prescriptive dimension of disagreement is only one element that makes iv) highly plausible. If iv) turned out to be false, then the here espoused line of argument should be discarded. Whether our notion of disagreement should be revised in such a deep way would be a far-reaching consequence of the issues raised in this paper. It should be conceded, nonetheless, that the charge of proof is so far clearly on the side of the objector that denies iv).

## 5 Kripke's real puzzle

Although I have called the strategy here criticized a 'Fregean' strategy, it is quite on the cards that the main problem these considerations raise has nothing particular to do with Fregeanism. Indeed, the key point of the foregoing discussion is probably, and perhaps ironically, in accordance with Frege's own views against the subjectivity of sense. For instance, in the final lines of a letter to Jourdain, Frege writes:

[10] Following Stevenson's lead, for instance, Ridge 2013 has defended that a satisfactory notion of disagreement must accommodate its fundamental prescriptive nature in terms of incompatible advices (and not simply in terms of incompatible beliefs or attitudes).

> 'Now if the sense of a name was something subjective, then the sense of the proposition in which the name occurs, and hence the thought, would also be something subjective, and the thought one man connects with this proposition would be different from the thought another man connects with it; a common store of thoughts, a common science would be impossible. It would be impossible for something one man said to contradict what another man said, because the two would not express the same thought at all, but each his own.' (Frege 1980: 80)

We may thus put the main conclusion of this paper in terms of Frege's remarks: if senses or analogous intensional categories are subjective (in the way standard solutions to Kripke's puzzle seem to require), then we lose sight of a 'common store of thoughts' and hence the possibility of contradiction at the level of thought necessary for disagreement.

This is not the place to elaborate this Fregean line of reasoning further. If the above considerations are sound, however, Kripke's puzzle is clearly a puzzle about rational *and* public belief. In particular, it seems of no use to save Pierre's rationality if the resulting notion of rationality does not allow for public disagreement/agreement phenomena compatible with Pierre's beliefs. The puzzle cannot be properly addressed without paying due attention to the tight connection between rationality and these public phenomena. Differing senses or modes of presentation are not, without further ado, compatible with elementary aspects of language communication of the sort Pierre's situation clearly requires.[11]

Víctor M. Verdejo
Departamento de Lóxica e Filosofía Moral
Facultade de Filosofía
Praza de Mazarelos s/n
15782 - Santiago de Compostela
University of Santiago de Compostela
victormartin.verdejo@usc.es

[11] Acknowledgements: I would like to thank an anonymous referee for Disputatio for insightful comments on an earlier version of this paper. This work has received financial support from the Spanish Ministerio de Economía y Competitividad (research projects FFI2009-08828/FISO and FFI2012-35153) and from the Catalan government, via the consolidated research group GRECC (SGR2009-1528).

## *References*

Chalmers, David. 2011. Propositions and attitude ascriptions: a Fregean account. *Noûs* 45: 595-639.

Crimmins, Mark. 1992. *Talk about Beliefs*. Cambridge, MA: MIT Press.

Feldman, Richard. 2006. Epistemological puzzles about disagreement. In *Epistemology Futures*, ed. by Stephen Hetherington, 216-236. New York: OUP.

Feldman, Richard and Warfield, Ted A. (eds.). 2010. *Disagreement*. New York: Oxford University Press.

Fine, Kit. 2007. *Semantic Relationism*. Oxford: Blackwell.

Forbes, Graeme. 1987. A dichotomy sustained. *Philosophical Studies* 51: 187-211.

Frege, Gottlob. 1980. *Philosophical and Mathematical Correspondence.* Edited by Gottfried Gabriel, Hans Hermes, Friedrich Kambartel, Christian Thiel, Albert Veraart and Brian McGuinness and translated by Hans Kaal. Oxford: Basil Blackwell.

Goldstein, Laurence. 2009. Pierre and circumspection in belief-formation. *Analysis* 69: 653-55.

Heck, Richard. 1995. The sense of communication. *Mind* 104: 79-106.

Kripke, Saul. 1979. A puzzle about belief. In *Meaning and Use*, ed. by Avishai Margalit, 239-83. Dordrecht: Reidel.

MacFarlane, John. 2007. Relativism and disagreement. *Philosophical Studies* 132: 17-31.

Richard, Mark. 1990. *Propositional Attitudes: An Essay on Thoughts and How We Ascribe Them.* Cambridge: Cambridge University Press.

Richard, Mark. 2011. Relativistic content and disagreement. *Philosophical Studies* 156: 421-31.

Ridge, Mike. 2013. Disagreement. *Philosophy and Phenomenological Research* 86: 41-63.

Salmon, Nathan. 1986. *Frege's Puzzle*. Cambridge, MA: MIT Press.

Schiffer, Stephen. 1992. Belief ascription. *Journal of Philosophy* 89: 499-521.

Sosa, David. 1996. The import of the puzzle about belief. *Philosophical Review* 105: 373-402.

Sundell, Timothy. 2011. Disagreements about taste. *Philosophical Studies* 155: 267-288.

Zalta, Edward. 1988. *Intensional Logic and the Metaphysics of Intentionality.* Cambridge, MA: MIT/Bradford.

Zalta, Edward. 1989. Singular propositions, abstract constituents, and propositional attitudes. In *Themes from Kaplan*, ed. by Joseph Almog, John Perry, and Howard Wettstein, 455-78. Oxford: OUP.

# Panpsychism without Subjectivity? A Brief Commentary on Sam Coleman's 'Mental Chemistry' and 'The Real Combination Problem'

**Michael Blamauer**
University of Vienna

BIBLID [0873-626X (2013) 37; pp. 299-309]

In a review, Sam Coleman 2012a praised Panpsychism as 'hot stuff' and I agree with him, because Panpsychism offers a theoretically elegant (even if somehow radical) way of handling the hard problem of consciousness within a moderate physicalist image of the world. If one considers experience as a fundamental property on a par with fundamental physical properties, then there are only two theoretical options: Either experience is a strongly emergent property of certain complex structures or it is ubiquitous.[1] So if one wishes to avoid dealing with the problem of how the experiential magically emerges from the non-experiential, Panpsychism seems to be the only option. However, with Panpsychism, philosophers can easily get their fingers burnt by touching on the Combination Problem — as does Coleman himself in his attempt to solve it.

Opponents of Panpsychism present the Combination Problem as quite comparable to the problem of strong emergence. While one

[1] I consider Panpsychism or Panexperientialism to be a theory that claims the ubiquity of mental or experiential properties, respectively. Thus, I consider Panpsychism or Panexperientialism not to be about proto-mental or proto-experiential properties. This is mostly because I don't see much explanatory power in these notions: If proto-mental properties are not mental, then they are physical and the hard problem returns with full force. And if proto-mental properties are in some sense mental, then the concept is delusive because it merely conceals problems with fundamental subjectivity such as those faced in the Combination Problem. In what follows, I will show that Coleman's position faces similar problems.

*Disputatio*, Vol. V, No. 37, November 2013
Received: 05/11/2012 Revised: 27/02/2013 Accepted: 30/05/2013

may have a hard time trying to understand how and why certain complex physical structures suddenly give rise to conscious experience, there also seems to be no easy answer to the question of how a certain number of lower single states of subjective experience can be combined to result in a unified higher (and qualitatively new) state of consciousness. Coleman correctly emphasized these two essential points in his papers 2012b, 2013: (1) The Combination Problem has its origins in the notion of fundamental subjectivity, and (2) without its solution, panpsychism loses most if not all of its explanatory power. However, the solution Coleman offers in his articles is comparable to cutting the Gordian knot: If the impossibility of a 'real combination'[2] of subjective simples lies at the heart of the Combination Problem, then the "essential part" of its solution is the 'disposal' of the notion of subjectivity on the fundamental level (2012b: 156). Now, having transformed subjectivity from a fundamental into a derivative, 'structural' feature (2013: 21) of certain organisms, nothing stands in the way of 'real combination' and the success of Panpsychism — or so Coleman claims.

In what follows, I will challenge Coleman's attempt to solve the Combination Problem in two steps. In section one (I) I will provide a brief sketch of Coleman's position which I will conclude by formulating three suspicions: (1) Coleman's approach to solving the Combination Problem by removing subjectivity from the fundamental level and transforming it into a derivative feature moves his own position close to a reductive representationalist account of consciousness or (2) moves it close to an emergentist account of consciousness (both of which stand in opposition to Panpsychism[3]); and (3) Given his reductionist account of subjectivity, he also cannot adequately solve

[2] For Coleman, 'real' combination is different to mere aggregation because it gives rise to a unified whole. And it is also different from a kind of combinatorial infusion such as Seager 2010 proposes, since the combinatorial parts do not lose their identity in favour of the emergent whole. For detailed discussion see Coleman 2013, section 7.

[3] As Coleman himself states: 'Panpsychism [...] stands opposed to emergentism.' (2013: 5) And further: 'For how could the conscious, the felt, the sentient, derive from the dead, the unfeeling, the insentient? That is why we have an explanatory gap and why, so say the panpsychists, conventional physicalism [i.e. reductionism about consciousness, MB] should be abandoned [...].' (2012b: 137)

the Combination Problem. In the subsequent section two (II) I will argumentatively flesh out these three suspicions to finally conclude that Coleman's approach to solving the Combination Problem fails for the given reasons.

## I

The Panpsychist's Combination Problem is, roughly put, the name for the fact that we currently have no clue of how a combination of micro-experiences may result in full-blown conscious experience like ours. The problem is rooted — so the suspicion — in the impossibility of subjects summing. This means that even if we could make sense of the idea of physical ultimates having conscious experience, there nevertheless seems to be no easy answer to the question of how a combination of these subjective ultimates may result in a qualitatively new 'higher' subject (of the kind we take ourselves to be): Combinations of micro-subjects simply do not seem to entail the existence of macro-subjects.

Now, Coleman correctly locates the most problematic assumption (which he tags the 'First Assumption', 2012b: 148, 154, 156) at work 'behind' the Combination Problem in the postulation that 'phenomenal ultimates are themselves subjects of experience' (2012b: 144). This assumption is a result of what Coleman calls the 'Quick Argument' (2012b: 148f), which states that experience requires a subject of experience that lives through it. If there is something it is like to experience, then there is a subjective point of view for which it is like. According to Coleman this is an 'apparent truism' for which no philosopher[4] has ever offered a convincing argument:

> 'The Quick Argument proceeds from a natural claim concerning phenomenal qualities, namely that where they exist they must be experienced by some subject. 'There cannot be experience without an experiencer', it is said. The next step is simply to apply this apparent truism to the panpsychist's ultimates.' (Coleman 2012b: 148, see also formulations on 152 and 153)

Even though one gets the impression that he somehow conceptually confuses the notions of phenomenal *properties* and phenomenal

[4] Coleman references Strawson 2003 in footnote 18 as a current defender of this claim.

*qualities*,[5] the stated goal of Coleman's critique of the 'First Premise' is the conceptual separation of two fundamental aspects of experience: (a) the phenomenal quality of an experience (or its 'phenomenal character') and (b) its being-for-a-subject-of-experience (or its 'subjective character').[6] Now, if the impossibility of a real combination of subjective simples lies at the heart of the Combination Problem — as Coleman compellingly argues to the reader —, then a possible solution might be found, he suggests, in simply taking phenomenal qualities of experience as an intrinsic, fundamental, feature of ultimates, and transforming the subjective character of experience (its being-for-a-subject-of-experience) into a derivative, structural, representational feature of certain macro-experiential systems. In short: For the purpose of solving the Combination Problem, Coleman suggests assuming that the intrinsic nature of the panpsychist's ultimates is devoid of any subjective character, and he instead favours 'phenomenal qualities'. If we examine Coleman's positive characterization of phenomenal qualities, we see that they derive exclusively from the sensory qualities present in our everyday experiences. His examples are taken mostly from vision (red London bus) or flavour (lasagne, Sunday roast beef), even though he assures the reader that 'phenomenal colours [...] are really just analogues to the true phenomenal natures of the ultimates.' (2012b: 155) However, the essential feature of phenomenal qualities in terms of Coleman's goal is explained negatively: Phenomenal qualities do not have to (necessarily) manifest themselves in experience in order to *be*, but can exist without being experienced. Phenomenal *qualities* are — unlike phenomenal *properties* — 'properties of objects' in the first place. (2012b: 150) The

[5] The claim that some special kind of qualities (namely phenomenal ones) cannot exist independently of a subject experiencing them is definitively something other than the claim that experience involves a subject of experience. The second claim is not about any kind of contents of my experience, but about the being of experience itself: it is about the question of what it is like to experience, hence about how my experience is in and as itself *for me*. To assume that some sort of *qualities* can exist *without being experienced by* a subject is different from assuming that *experience* can exist without there *being a subject for whom it is like* to do so. The first assumption concerns phenomenal *qualities*, the second phenomenal *properties*.

[6] The conceptual distinction between 'phenomenal character' and 'subjective character' and their relationship is taken from Kriegel 2009 and 2011.

essential upshot of his explanation is that phenomenal qualities can be treated as objective qualities. Of course, if we then take 'the question of how phenomenally-qualitied items combine' (2012b: 138) as an expression of the Combination Problem of Panpsychism, then, I suspect, we have found a reason for why Coleman presents his alleged solution in terms of cooking or painting (for detailed examples see 2012b: 140 and 157f).

However, given that the real combination of phenomenal qualities is unproblematic, Coleman still needs to say something about the subjective character of consciousness, its being-for-a-subject-of-experience: 'We [...] need to say something about how genuine subjects, beings like ourselves, arise on the present picture.' (2012b: 154) Here, Coleman makes an interesting move:

> '[S]ince subjects cannot combine into larger subjects, the only way to preserve the panpsychist anti-emergence principle when it comes to high-level subjecthood is to allow that, while quality is a fundamental affair, subjectivity must be susceptible of a reductive treatment.' ( Coleman 2013: 21)

And further:

> 'Conscious awareness as we know it is therefore to be thought of as *phenomenal representation*, the representation of phenomenal quality by phenomenal quality.' (Coleman 2012b: 159)

It is here that I detect Coleman's most problematic claim, a claim that will finally lead — as I will show — to the collapse of his whole position and its central attempt to solve the Combination Problem. In the subsequent section I will dispute his reductionist claim about subjectivity by fleshing out the following three suspicions: (1) It is impossible to reduce subjective experience to a-subjective qualities and their representational relations without claiming that subjectivity is illusory (which I find deeply counterintuitive); (2) Every attempt of retaining the fundamentality of subjectivity within Coleman's theoretical framework would require the notion of emergence (which runs counter to one of Panpsychism's central premises) and (3) Even if we assume for the sake of argument that Coleman's approach is sound, the Combination Problem would nevertheless remain unsolved in that it is perfectly conceivable (and therefore possible) that a representational state of phenomenal qualities exists without there being consciousness (in the sense we are acquainted with). To this end, I begin with a brief sketch of Coleman's own 'positive' representational account of subjectivity.

## II

As noted, Coleman's key move to solve the Panpsychist's Combination Problem consists in removing subjectivity from the fundamental, i.e. constitutive level, and transforming it into a derivative structural feature: 'Panpsychists hold, effectively, that all non-fundamental properties are structural: they are reducible to more basic properties plus arrangement of their bearers. That is the non-emergence principle.' (2013: 21) Thus, 'subjectivity must be susceptible of a reductive treatment.' And he suggests the following reductive picture of subjectival awareness:

> 'It is [...] the essentially *structured* (composite) nature of the phenomenally-qualitied systems posited that enables them to be subjects of their own phenomenal qualities [...]. [...] Conscious awareness as we know it is therefore to be thought of as *phenomenal representation*, the representation of phenomenal quality by phenomenal quality.' (Coleman 2012b: 159)

And further:

> 'To be such a representational system is to be conscious in the way that we recognize each in our own case.' (Coleman 2012b: 160)

Now, to see the problem with this account it is necessary to first say something about the subjective nature of experience. I think I am not alone by holding to the claim that consciousness (or experience in the way we are acquainted with in everyday life) necessarily entails a subject of experience for whom it is somehow or other like to have this experience: If there is something it is like to be in a state of pain, then, necessarily, there is something it is like *for* someone or something to be so. Regarding Coleman's separation of the two central aspects of experience — phenomenal quality and subjectivity — it is important to emphasize that, whereas the phenomenal quality of an experience characterizes it just as the experience it actually is (in contrast to qualitatively different experiences), its subjective character (or the being-for-a-subject-of-experience of these qualities) is what makes the experience an experience at all: The feeling of pain is qualitatively different from the sensations you get while swallowing chocolate, but what sense does the difference make if there is no one (no subject of experience) for whom the difference is manifest because there is nobody who actually feels the pain or tastes the chocolate?

Now, an essential feature of subjective experiences in general is their indubitability due to our direct acquaintance with them. This special kind of direct acquaintance of conscious experiences with themselves (call it primitive self-consciousness or subjectivity) is the reason for the distinction between a first-person and a third-person ontology, which provides the basis of most of the arguments against materialism, like the zombie-argument or the explanatory-gap argument. For example, when I taste a cold, clear and transparent liquid, thinking it is water and being told afterwards that it is not $H_2O$ but XYZ, I was wrong about the *content's* being (in this case, that I tasted water). But I definitely was not wrong about the *experience's* being: I tasted a cold, clear and transparent liquid because I was immediately and *indubitably aware* of it by having the experience. This is even more obvious in cases where the content's being is strictly tied to its appearing in experience. Think of pain: what *appears* be pain to a subject *is* pain, because this subject cannot coherently deny its existence due to its immediate and indubitable presence in her experience. And vice versa, if she *denies feeling* pain, then she *denies* there *being* pain due to its indubitable absence in her experience. Thus there can be no doubt about the existence of an experience from a first-person perspective, irrespective of all other physical facts.

In the following I will show that this special character of consciousness poses a real problem for Coleman's representationalist account of subjectivity. Following Coleman, subjective awareness, i.e. awareness of phenomenal qualities, is an extrinsic, additional feature to the specific quality an experience might have. Combined with his 'non-emergence principle', which states that all structural properties are principally reducible to more basic properties and their relations, this implies that subjectival awareness logically supervenes on the complex arrangement of fundamental simples and their intrinsic a-subjective phenomenal qualities. This opens three ways of criticism.

The first is based on some strong intuitions we have about our everyday conscious lives and which relate to what has been said about our direct acquaintance with them. Given the correctness of the presented assumptions about the immediacy and indubitability of the givenness of experiences for a subject of experience, conscious subjects who feel pain can never be wrong about themselves being in a state of pain, irrespective of whether the underlying representational structure implies the contrary. And the same is true vice versa. Subjectivity is simply not identical with this representational structure. Yet Coleman's strong reductionist claim about subjectival awareness being a structural feature seems to imply precisely this. There can never be a case where ultimates stand in the right relations to each other but fail to instantiate a point of view, i.e. a subject of experience, because in the reductionist account subjectival awareness follows logically from the underlying representational state (due to their identity). But this runs contrary to our basic intuitions about the being of our experiences. For example, it is perfectly conceivable (and, I assume, therefore possible) that subjectival awareness exists (I am in pain and I am certain of this fact) without the realization of the adequate representational state of phenomenal qualities. Or, vice versa, you can also never rule out a state that represents oneself as a conscious system (probably producing a verbal output like: 'I am conscious of your delicious lasagne'), but does not actually instantiate subjectival awareness (i.e. that there is nothing it is like to be me because there is no 'me'). I think the common antireductionist arguments (the explanatory gap, the possibility of zombies etc.) provide a fairly good basis for considering the subjectivity of consciousness (the one we are directly acquainted with in everyday life) as a fundamental feature, because it is the feature which distinguishes consciousness — as a phenomenon with a first-person ontology — from third-person extrinsic phenomena, which furthermore is the reason why it is neither reducible nor adequately explainable in terms of (third-person, extrinsic) structures and relations (this is what the hard problem is all about).

But if subjectivity — as I have tried to argue — is not only an essential but a fundamental aspect of consciousness, then the only way to save Coleman's position from the aforementioned problems (in fact, problems all reductionist accounts of consciousness face) would be to claim that subjectivity is an irreducible feature — yet not a feature of fundamental ultimates.

This brings me to the second aspect of my critique, which can be put rather briefly: Given we understand emergence as the sudden coming-into-existence of ontologically new properties,[7] which preclude reduction to more basic ones, then subjectivity — by taking it as irreducible — would turn into an emergent property. But this option definitely is a no-go — and I assume Coleman would agree — in that it questions the whole project of Panpsychism, which is essentially based on the idea of smooth evolution and the lack of emergence.

Thirdly and finally — considering all that has been said — let us take another look at the Combination Problem within Coleman's framework. This third and last aspect of my critique is based — like the first one — on conceivability issues regarding micro- and macro-phenomena.[8] Coleman's Panpsychism (or rather Pan-proto-psychism because he only claims the ubiquity of phenomenal *qualities*, not phenomenal *properties*) is a form of *constitutive* Panpsychism, which means that the micro-level facts (about phenomenally-qualitied ultimates and their relations) constitute the macro-level facts (about conscious experience as we know it). As Coleman has convincingly argued, if constitutive Panpsychism claims the ubiquity of *experience* (and thus the ubiquity of subjectivity), then it faces the Combination Problem, because the micro-level facts about experiential ultimates (and their relations) do not entail the macro-level facts about consciousness: A set of micro-subjects simply cannot really combine to constitute a further, qualitatively new subject. Now, let us turn to Coleman's proposed solution to the problem.

If ultimates are no longer bearers of phenomenal *properties* (experience, subjectivity), but merely phenomenal *qualities*, then 'real combination' seems to be no big deal. And, further, if subjectivity is taken as a reductive feature, logically supervening on the representational states of the phenomenally qualitied ultimates, then subjective

[7] Or, in Coleman's own words: 'paradigm cases of emergence' are 'cases where the underlying properties cannot generate their product structurally, because inputting *what they are* [...] makes no contribution towards what results.' (2013: 18)

[8] Here, I am mainly following Goff 2009 and his concept of panpsychist's zombies.

consciousness seems to be entailed by the facts about the phenomenally-qualitied ultimates and their representational relations. Thus, given all the micro-facts, the macro-facts about consciousness come 'free of charge'. But could this really be seen as a solution to the Combination Problem? I thoroughly doubt it, for reasons presented already in the first aspect of my critique: Given all the micro-level facts about qualitied ultimates and their relations, it is nevertheless perfectly well conceivable (and therefore, as I assume, possible) that even though all the qualitative aspects of an experience (on the macro-level) are instantiated, subjectival awareness of the qualities is not, because subjectival awareness is not entailed by the micro-level facts. But if subjectival awareness is not instantiated, then neither is consciousness. The result would be a kind of panpsychic zombie (see Goff 2009), a being qualitatively identical to me, but lacking consciousness, because there would be no 'me' — so to speak — for whom it could be like to have those qualities present.

If we assume that the essential goal of Panpsychism is to make sense not just of the existence of phenomenal qualities manifest in experience, but also of the existence of an experiential, subjective, first-person perspective in the first place, then Coleman's attempt to solve the Combination Problem with his (to my mind oxymoronic) version of an a-subjective, constitutive Panpsychism simply fails.

## Conclusion

Coleman's attempt to solve the Combination Problem fails for the reasons stated above. However, despite the Combination Problem, Panpsychism seems to remain a viable candidate for a theory of consciousness, since it attempts to apply fundamental subjectivity to a moderate physicalist view of the world. Coleman writes:

> 'We really should want to say something remotely interesting about how minds come about, not simply take them so thoroughly for granted. It is not just that this position is implausible, it is that solving practically any problem in this way is fundamentally boring.' (Coleman 2012b: 149)

But perhaps he is being too sensationalist. With respect to the Combination Problem, I do not see why Panpsychism should be 'fundamentally boring': The challenge lies precisely in searching for a framework that handles the tension between fundamental subjectivity and objectivity, unity and diversity and the question of real combination. Coleman merely attempts to resolve this tension via a reductionist move and I do not see why that should be more interesting.[9]

Michael Blamauer
Department of Philosophy
University of Vienna
Universitätsstraße 7, 1010
Vienna, Austria
michael.blamauer@univie.ac.at

## *References*

Coleman, Sam. 2012a. Review of 'The Mental as Fundamental' Ed. Michael Blamauer. *Notre Dame Philosophical Reviews*.

Coleman, Sam. 2012b. Mental Chemistry: Combination for Panpsychists. *dialectica* 66 (1): 137–166.

Coleman, Sam. 2013. The Real Combination Problem: Panpsychism, Micro-Subjects and Emergence. *Erkenntnis* (DOI 10.1007/s10670-013-9431-x)

Goff, Philip. 2009. Why Panpsychism Doesn't Help Us Explain Consciousness. *dialectica*, 63 (3): 289–311.

Kriegel, Uriah. 2009. Self-Representationalism and Phenomenology. *Philosophical Studies* 143: 357-381.

Kriegel, Uriah. 2011. Self-Representationalism and the Explanatory Gap. In *Consciousness and the Self: New Essays*. Cambridge: Cambridge University Press.

Seager, William. 2010. Panpsychism, Aggregation and Combinatorial Infusion. *Mind & Matter* Vol. 8(2): 167-184.

Strawson, Galen. 2003. What is the Relation between an Experience, the Subject of the Experience and the Content of the Experience?. *Philosophical Issues* 13 (1): 279–315.

[9] I would like to thank Wolfgang Fasching and the two anonymous referees of *Disputatio* for their helpful comments on a former draft of this paper.

*Disputatio*'s Symposium on Berit Brogaard's *Transient Truths*
Oxford University Press, 2012

# Précis

**Berit Brogaard**
University of Missouri-St. Louis

BIBLID [0873-626X (2013) 37; pp. 311-314]

*Transient Truths* is a part of a larger philosophical project that I have been interested in since I first started thinking about philosophical issues relating to the reality of time and tense. One issue having to do with the reality of time and tense is metaphysical. Some hold that tense is a feature of language but not of propositions, mental content or the world. On this B-theoretical view, the present moment is so-called, not because it is special, but because we perceptually experience only present entities. Others think that tense can be a feature of all of these entities. If the world is tensed, then the present has a different ontological status than the past and the future. I fall into the latter camp of 'serious tensers'. In previous work I have defended presentism, a form of serious tensism that implies that only present entities exist (e.g. Brogaard 2013a).

The main goal of *Transient Truths*, however, was not to defend the view that the present moment is special but to provide a book-length defense of a particular theory of propositions known as 'temporalism'. To a first approximation, temporalism is the view that there are propositions that can change their truth-values across time. There is no straightforward argumentative route from this view to A-theoretical views about time and tense. In fact, as I argue in Chapter 7, there is some reason to think that a B-theorist cannot adequately express her views if she rejects the temporalist approach to language. Although temporalism does not imply that the A-theory is correct, I do think that the debate has potential metaphysical implications. (Semantic) eternalism, the opponent view to the effect that all propositions have their truth-values eternally, together with some widely held assumptions, appears to have B-theoretical implications. One argument may run as follows. Propositions that are eternally true are not tensed. For any true fact there is at least one corresponding true

Received: 03/10/2013 Accepted: 09/11/2013

proposition that correctly represents that fact. As eternalism holds that there are no tensed propositions, the world cannot be tensed if eternalism is true. So, if the A-theory is true, then temporalism is true.

Temporalism is also of interest to me on grounds that are independent of metaphysics. I found the view intuitively appealing long before I started working on the book. But as I was exploring the literature I soon realized that whereas eternalism was widely held to be true by a long list of philosophers (e.g. Frege 1979; Stalnaker 1970; Lewis 1980; Richard 1981), temporalism was a minority position defended only by a few authors (Prior 1957, 1959, 1967, Kaplan 1989, Ludlow 1999, among others). This was the ultimate factor motivating me to write a book-length defense of the position.

The argumentative strategy of the book is to provide a functional account of propositions and then show that temporal content can play the functional role. Propositions are standardly held to be the semantic values of truth-evaluable sentences, the object of propositional attitudes, the objects of agreement and disagreement, the contents that are passed on in successful communication, and the contents that intensional operators operate on. On the functional approach, entities that best satisfy these descriptions count as propositions. In the book I present a wide range of arguments for believing that temporal propositions can play this role and reply to a number of traditional arguments for thinking that they do not function in this way. Since temporalism does not say that all propositions are temporal, showing that temporal contents sometimes play the proposition role suffices to establish the truth of the doctrine.

Though there are authors who have argued that all propositions are temporal, I offer some reasons in Chapter 7 for thinking that this is not so. While temporal contents can, and often do, play the role of propositions, eternal contents can also play this role. So, on the view I defend, there are both eternal and temporal propositions. Eternal propositions are, for example, expressed by language that serves the purpose of describing metaphysical positions. For instance, we cannot confirm or deny the view that only present things exist or that only present events are happening without using language that expresses tenseless propositions. However, there is no reason to think that this type of language, even if true, is made true by tenseless facts in the world. So, while I think that some form of the correspondence theory of truth is correct, I reject the traditional, structural correspondence theory, according to which true propositions completely mirror reality.

A word about the book's structure: After clarifying some conceptual issues in the book's first chapter I argue that temporal contents are the main objects of belief and other propositional attitudes in Chapter 2. In Chapter 3 I offer arguments for the view that temporal contents are the main objects of agreement and disagreement. The two subsequent chapters argue that the eternalism/temporalism debate is directly related to the debate about whether the tenses function as sentential 'index-shifting' operators, and I provide an outline of a operator theory of tense. In chapter 6 I argue against a version of eternalism that grants that tense operators operate on temporal contents but denies that temporal contents are propositions. I then consider the question of whether there are eternal propositions. In the final chapter I extend some of the considerations of the previous chapters to the case of perceptual experience.

Though a lot can be said in the course of a whole book, there is much more to be said about these issues than I was able to fit in. I am happy to have the opportunity here to engage with three bright thinkers in further debate about these issues. The points they bring up contribute in significant ways to the debate about eternalism and temporalism as well as the larger picture about the metaphysics of time.

Berit Brogaard
University of Missouri-St. Louis
Department of Philosophy
599 Lucas
1 University Boulevard
St. Louis, MO 63121-4499
314-516-5631
brogaardb@umsl.edu

## *References*

Frege, G. 1979. Logic. In *Posthumous Writings,* ed. by H. Hermes, H. Kambartel, and F. Kaulbach, 126-51. Tr. by Long and White. Chicago: Chicago University Press.

Kaplan, D. 1989. Demonstratives. In *Themes from Kaplan*, ed. by J. Almog, J.

Perry and H. Wettstein, 481-563. New York: Oxford University Press.

Lewis, D. 1980. Index, context and content. In *Philosophy and Grammar*, ed. by S. Kanger and S.Ohman. Reidel, Dordrecht, 79–100. Reprinted in *Papers on Philosophical Logic*, 21-44. Cambridge University Press, 1998.

Ludlow, P. 1999. *Semantics, tense, and time: An essay in the metaphysics of natural language*. Cambridge, MA: MIT Press.

Prior, A. N. 1957. *Time and Modality*, Oxford: Oxford University Press.

Prior, A. N. 1959. Thank Goodness that's over. *Philosophy* 34: 12-17.

Prior, A. N. 1967. *Past, Present and Future*, Oxford: Clarendon.

Richard, M. 1981. Temporalism and Eternalism. *Philosophical Studies* 39: 1-13.

Stalnaker, R. 1970. Pragmatics. Synthese 22. Reprinted in Stalnaker 1999.

Stalnaker, R. 1999. *Context and Content*, Oxford: Oxford University Press.

*Disputatio*'s Symposium on Berit Brogaard's *Transient Truths*
Oxford University Press, 2012

# Propositions and the Metaphysics of Time

**Giuliano Torrengo**
University of Milan

BIBLID [0873-626X (2013) 37; pp. 315-321]

The central point of Brogaard's interesting essay is that *temporalism*, roughly, the thesis that there are propositions whose evaluation is sensitive to time (14), is a better alternative to standard *semantic eternalism*, roughly, the thesis that no proposition is sensitive to temporal variation. Five theoretical roles individuate propositions: (i) semantic values of sentences, (ii) objects of attitudes, (iii) objects of agreement and disagreement, (iv) what is transferred in successful communication, and (v) what intensional operators operate on (5-6, 30). For each of those roles Brogaard aims at showing that *temporal* propositions fare better than eternal ones, in that the problems that have been traditionally raised against them can be overcome, while more serious problems can be raised against eternal propositions. Although most of the book touches upon issues of philosophy of language and philosophy of mind, as it should be, there are several considerations that Brogaard makes about the relation between the semantic tenets of temporalism and eternalism, on the one side, and the metaphysics of time, on the other. In this note, I will assume that the semantic arguments in the book for preferring temporalism over eternalism are sound, and confine myself to some criticism concerning their metaphysical import.

In the introduction, Brogaard points out that semantic eternalism is metaphysically more demanding than temporalism, since it entails *metaphysical eternalism*, namely the view that past and future times are ontologically on a par with the present — or at least because it rules out *presentism*, the thesis that only present entities exist (6-7). The reason is that eternal propositions can contain non-present times as constituents. Assuming that the existence of a proposition entails the

*Disputatio*, Vol. V, No. 37, November 2013
Received: 03/10/2013 Accepted: 09/11/2013

existence of its constituents, it follows that eternal propositions containing reference to non-present times are incompatible with presentism. Of course, a presentist could think of times as some sort of *ersatz entities*, namely sets of temporal propositions. But the propositions that individuate times are temporal propositions, and thus they are not at the semantic eternalist's disposal (see 7, note 4).

I agree with Brogaard's remarks. Indeed, while there is no immediate objection to the use of ersatz times as parameters of the circumstances of evaluation that the presentist makes, appealing to a set of propositions as a constituent of a proposition seems suspiciously circular. However, I do not think that those arguments can be used to support temporalism, at least the variety of temporalism defended in the book. Here is the problem: semantic eternalism is supposed to put stricter constraints on the choice of a metaphysics of time than temporalism. However, it is the thesis that *some* eternal propositions exist that is at odds with presentism, rather than the stronger thesis that *all* propositions are eternal. And many temporalists do admit that not all propositions are temporal. Brogaard herself defends a version of temporalism in which eternal propositions that make reference to non-present times are accepted (155-57). Maybe I am wrong and presentism is compatible after all with propositions that make explicit reference to a non-present time. But if I am wrong, because presentist can resort to ersatz times as constituents of propositions, say, then it seems that semantic eternalism is compatible with presentism after all. I can imagine an 'in between position' to the effect that the temporalist can accept ersatz times as constituents of eternal propositions, while the eternalist cannot. But I do not think this view is very appealing, and it is defended nowhere in the book.

Be that as it may, I have another concern with respect to how Brogaard characterizes the relation between semantic eternalism and metaphysical eternalism. According to her: 'semantic eternalism [...] makes it difficult for metaphysical eternalists to articulate the commitment of their theories' (7). This thesis is first stated in the introduction and then elaborated at length in Chapter 7. Here is what I take to be the core's of Brogaard's position. Presentists and eternalist are taken to disagree on whether wholly past objects, such as Socrates, exist. Thus, a way to state their disagreement is to say that they disagree on how (2) below [Chapter 7's numeration] is evaluated when uttered, say, in 2013.

(2) Socrates exists.

However, ordinarily understood, (2) is taken to convey that Socrates is located roughly in the same spatiotemporal location in which we are, and thus we can meet him as we meet our friends. Both the eternalist and the presentist agree on the falsity of this reading. In the literature, many agree that in order to avoid a skeptical outcome with respect to the substantivity of the eternalism/presentism distinction, something like such a 'philosophy room reading' of (2) must be at both the eternalist's and the presentist's disposal[1]. Roughly the idea is that what the eternalist means with (2), when she disagrees on the truth value of an utterance of (2) in 2013 with the presentist, is that quantifying *unrestrictedly* — in particular, without any restriction on the temporal dimension — (2) expresses a true proposition.

Now, why, according to Brogaard, is the metaphysical eternalist who also endorses semantic eternalism in trouble? The idea is that from eternalism it follows that any tensed sentence expresses an eternal proposition relative to its context of utterance. Thus, (2) as uttered at time $t*$ expresses the proposition that is more perspicuously expressed by (2a) below.

(2a) Socrates exists at t*.

However, (2a) seems to be as ambiguous as (2), and thus it does not qualify as a good *analysis* of the 'philosophy room' reading of (2). Providing that a minimal condition on a good analysis is to yield an unambiguous paraphrase. Brogaard's proposal is to endorse temporalism instead, and to allow for two readings of (2) in terms of the proposition that they express. According to the first reading — the ordinary reading — (2) expresses a temporal proposition that gets evaluated at the time of utterance. In this reading, both the presentist and the eternalist agree that (2) is false. According to the second

[1] See Zimmermann 1998, Oaklander 2002, Sider 2006, Tallant 2013 and the debate between Crisp 2004a, 2004b and Ludlow 2004. Meyer 2005, Savitt 2006, Dorato 2006, Callender 2011 hold a sceptical position.

reading — the philosophy room reading — (2) expresses an eternal proposition that only the eternalist accepts as true.

I have three qualms here, the first and the last with Brogaard's positive proposal, and the second with her criticism of eternalism's expressive capacity. As for the first qualm, according to Brogaard, and coherently with her view, the eternal proposition that *Socrates exists* is evaluable as true or false *simpliciter* only in a context in which either Socrates is an instantaneous object, or Socrates always (or never) exists (150). But then, if Socrates is not a instantaneous object, the presentist cannot claim that (2) is *false* in the philosophical room reading. She can disagree with the eternalist only in the sense that she does not accept (2) as true. Maybe that's enough for making the disagreement substantial, but if an alternative accounts in which (2) turns out false is available to the presentist, there seem to be reasons to prefer it. I understand that Brogaard endorses four-dimensionalism as a theory of persistence, and thus she does have an account in which the eternal reading of (2) is false *simpliciter*, since she maintains that in that reading (2) is about a instantaneous temporal slice of Socrates. However, dialectically, the fact that if one assumes temporalism, then one *must* endorse a particular thesis about persistence in order to express a distinction about temporal ontology seems to me a drawback of temporalism.

Secondly, it is not clear to me why the fact that a *sentence* such as (2a) is ambiguous between a restricted reading and an unrestricted reading of the quantifier counts as evidence against semantic eternalism. Brogaard's worry seems to be that the role of the time of utterance $t^*$ in the proposition expressed is unclear. But once we accept the distinction between a temporally restricted and a temporally unrestricted reading of quantification (and something analogous for predication), the worry is spurious. The role of the temporal parameter provided by the context within the content expressed depend on the particular form of semantic eternalism, but in no case it will determine a restriction on the quantifier *on an unrestricted reading of the quantifier*. Of course, if the quantifier is understood as temporally restricted, then the time of utterance together with tenses and possibly other (pragmatic and semantic) elements will determine the restriction. But if we agree that the distinction between temporally restricted and temporally unrestricted interpretation of the quantifier is understood, then Brogaard's complain vanishes.

Generally speaking, the problem of articulating the commitments of presentism and eternalism seems to me orthogonal to the debate on temporalism vs. eternalism. In order to state the disagreement between the presentist and the eternalist, we need to understand (2) in such a way that the extension of 'exist' is not limited by temporal factors[2]. Once we grasp this unrestricted construal of 'exist', if we are temporalists, we will take the claim to express a temporally neutral proposition to be evaluated relative to the time of utterance; if we are eternalist, we will take it to express a content that is indexed to the time of utterance, and thus temporally invariant. Maybe many metaphysical eternalists, after having read Brogaard's semantic arguments in favor of temporalism, will decide to endorse her view (I may be one of them), but semantic eternalism as such is not an obstacle to understanding temporally unrestricted quantification.

What would be an obstacle to articulate presentism and eternalism is ruling out *by semantics alone* the possibility that what exists *simpliciter*, namely unrestrictedly speaking, changes over time. The core of the distinction after all is that according to the eternalism what exists *simpliciter* never changes through time, whereas for the presentist it does. And it is important that such claims be stated by using the *same* language, which would be impossible if the language itself ruled out one of the two positions. But semantic eternalism does *not* rule out this possibility, in so far as it allows for tensed *sentences* to express claims of existence *simpliciter*. This leads me to my last worry.

When the eternalist refers to *presently* existing things, it seems natural to think that the agreement with the presentist is not confined only to the ordinary readings of the claims (on whose truth-value she agrees with the presentist also in the case of past entities), but it also reaches the unrestricted reading. But this does not seem to be the case if the unrestricted reading is interpreted as the reading that expresses an eternal proposition that does not contain a time constituent, as Brogaard maintains (148). The reason is that Brogaard is compelled to maintain that existential eternal propositions about non eternal entities are *never* true for the presentist, not even when uttered when the entity at issue is present. Therefore,

[2] I have argued that in Torrengo 2012.

there is a sense in which the presentist and the eternalist necessarily disagree on what *presently* exists, which seems to me a weird outcome of the view.

Imagine the following situation. A presentist and an eternalist find themselves in a biology lab. A scientist is observing through a microscope an amoeba (call it 'Amoeba'), which is about to undergo a process of a mitosis. Suppose that at *t*, Amoeba has not undergone mitosis, while at *t'* the process is complete, and assume that individual amoebas do not survive processes of mitosis (choose some other fatal event for protozoa otherwise). Consider now claim (1) below, and imagine that the eternalist utters it once at *t* and a second time at *t'*.

(1) Amoeba exists.

In the ordinary reading, the presentist and the eternalist agree that (1) is true when uttered the first time (at *t*), and false when uttered the second time (at *t'*). In the philosophy room reading, the two will disagree at *t'*: according to the eternalist, the eternal proposition that Amoeba exists is true at *t'*, whereas the same cannot be said of the presentist. But what about an utterance of (1) at *t* in the unrestricted reading? According to Brogaard's proposal, it should be taken to express a eternal proposition with no time index, which has the same truth vale with respect to *t* and *t'*. This is fine for the eternalist, who evaluates both utterances of the unrestricted reading as true. But it is bad news for the presentist. Since she does not accept (1) as true at *t'* in the unrestricted reading, she is compelled to do the same with an utterance of (1) at *t* — the eternal proposition expressed in both contexts being the same. Can't we say that the presentist at *t* agrees with the eternalist that Amoeba exists *simpliciter*? No, because, again, that would entail that Amoeba *will* exist *simpliciter* at *t'* too (160). Notice that according to the 'standard' semantic eternalist, utterances of tensed sentences such as (1) express *time indexed* eternal proposition. Thus, an utterance of the same *sentence* expresses different (time indexed) eternal propositions on different occasions, even when read unrestrictedly. This is good news, because it allows us to express the key difference between the presentist, who takes unrestricted existential claim to vary over time, and the eternalist, who does not. Finally, Brogaard argues convincingly that for the roles that propositions usually have in philosophy of language and mind the best option is often to endorse temporalism. However, when it comes to metaphysics, the traditional eternalist approach looks to me on more steady ground.

Giuliano Torrengo
Università degli Studi di Milano
Dipartimento di Filosofia
via Festa del Perdono, 7
20122 - Milano
Italy
giuliano.torrengo@unimi.it

## *References*

Callender, C. 2011. Time's Ontic Voltage. In *The Future of Philosophy of Time*, ed. by Adrian Bardon. London & New York, Routledge.

Crisp, T. 2004a. On Presentism and Triviality. In *Oxford Studies in Metaphysics*, ed. by D. Zimmerman, 1 15-20

Crisp, T. 2004b. Reply to Ludlow. In *Oxford Studies in Metaphysics*, ed. by D. Zimmerman, 1 37-46.

Dorato, M. 2006. The irrelevance of the presentist/eternalist debate for the ontology of Minkowski spacetime. In *The Ontology of Spacetime*, ed. by D. Dieks, 93-109. Elsevier.

Ludlow, L. 2004. Presentism, Triviality, and the Varieties of Tensism. In *Oxford Studies in Metaphysics*, ed. by D. Zimmerman, 1 21-36

Meyer, U. 2005. The Presentist's Dilemma. *Philosophical Studies*, 122: 213–225

Oaklander, N. 2002. McTaggart's Paradox Defended. *Metaphysica: International Journal of Ontology and Metaphysics*, 3, 1: 11-25

Savitt, S.F. 2006. Presentive and Eternalism in Perspective. In *The Ontology of Spacetime*, ed. by D. Dieks, Elsevier (http://philsci-archive.pitt.edu/archive/00001788/01/PEP.pdf)

Tallant, J.C. 2013. Defending Existence Presentism. Forthcoming in *Erkenntnis*

Sider, T. 2006. Quantifiers and Temporal Ontology. *Mind* 115 (457): 75-97

Torrengo, G. 2012. Time and Simple Existence. *Metaphysica*, 13: 125-130

Zimmerman, D. 1998. Temporary Intrinsics and Presentism. In *Metaphysics: The Big Questions*, ed. by P. van Inwagen and D. Zimmerman, 206–219. Malden (Mass.), Blackwell.

*Disputatio*'s Symposium on Berit Brogaard's *Transient Truths*
Oxford University Press, 2012

# Temporalism and Composite Tense Operators

**Dan Zeman**
Universitat Pompeu Fabra

BIBLID [0873-626X (2013) 37; pp. 323-328]

Berit Brogaard's book, *Transient Truths. An Essay in the Metaphysics of Propositions* is the most complete and thorough defence of temporalism to date. *Temporalism*, as she understands it, is the view that the objects of our attitudes such as belief, fear, desire and so on, as well as the entities expressed by some of our utterances, are temporal propositions — that is, contents that change their truth value over time. The opposite view, considered the orthodox view nowadays, is the view that such contents are eternal propositions — that is, contents that have their truth value eternally — hence, *eternalism*. In her book, Brogaard investigates all the major arguments against and in favour of temporalism, offering a couple of new ones in its favour along the way. One of the nice features of the book is that the interconnections between the various areas of philosophical inquiry in which temporalism is a competitor clearly come to the fore, while the contribution made by the view in each area is thoroughly investigated are carefully argued for. Thus, temporalism is put forward in connection to the theory of belief possession and retention (Chapter 2), to that of communication and of disagreement (Chapter 3), to the syntactic and semantic theory of tenses and temporal expressions (Chapters 4 to 6) and to the metaphysics of time (Chapters 7 and 8).

Given the book's wide range, there are many interesting and important issues to pick up. In this short note, I will focus only on the issue of the linguistic representation of tenses and temporal expressions, a field in which temporalism has been one of the traditional answers. This traditional answer, however, has been called into question in recent years, and a lively debate has issued. It is this debate that I want to focus on, and discuss Borgaard's contribution

*Disputatio*, Vol. V, No. 37, November 2013
Received: 03/10/2013 Accepted: 09/11/2013

to it. I will thus start with presenting the challenges she attempts to answer to, sketch her solution and then raise some worries about her proposal.

In chapter 4, Brogaard enters the debate between temporalism and eternalism in the context of providing the best syntactic and semantic theory of tenses, and purports to offer an answer to the often rehearsed arguments against temporalism summarized in King 2003, 2007. Those arguments show that the operator treatment to tenses and temporal expressions temporalists favour is, if not strictly speaking incorrect, more cumbersome, ad-hoc and significantly departing from the surface structure of English (King 2007: chapter 6). What King points out is that the operator treatment of tenses has at least prima facie problems with linguistic phenomena such as the interaction between tenses and temporal adverbials ('Yesterday, John turned off the stove'), temporal anaphora ('Sheila had a party last Friday, and Sam got drunk') and the so-called 'Kamp/Vlach sentences' ('One day, all persons alive now will be dead') — among others. Given that a quantificational treatment of tenses yields simpler results in dealing with these phenomena, and the (sociological) fact that the majority of linguists have given up the Priorian framework and have adopted a quantificational or referential approach to tenses instead — King argues —, temporalism should be abandoned.[1]

In response, Brogaard proposes a new view of tenses and temporal expressions that is supposed to vindicate temporalism and which also accounts for the problematic phenomena mentioned by King. First, Brogaard distinguishes between two kinds of tense operators: basic ones, such as 'It was the case that' and 'It will be the case that', and composite ones, which result from the combination of basic tense operators with a variety of temporal adverbial phrases. Within the last we find indexical frame adverbials ('this morning'), non-indexical frame adverbials ('in June 2030 (CST)'), durative adverbial phrases ('for two weeks'), mixed durative and frame adverbials ('for the last two hours'), adverbial subordinate clauses ('before I

[1] How strong King's actual claims are, and whether they licence the abandonment of the Priorean framework, is a matter of dispute. While it is true that most of those opposed to temporalism have taken King to offer a clear refutation of the view, not everybody is convinced — see Recanati 2007 and Marti & Zeman 2010, for example.

was born') and adverbial phrases of number and frequency ('never'). Although not all such adverbial phrases behave similarly in combining with tenses (for example, some of them 'scope out', while others don't), the combination of most of them with basic tense operators results in composite tense operators that are interpreted as circumstance-shifting sentential operators. Second, and this is the core claim of the view, the problematic phenomena mentioned by King can be accounted for by appeal to composite tense operators. In Chapter 4 Brogaard shows how this view accommodates the phenomena mentioned above, while in Chapter 5 she applies it to yet other troublesome phenomena, such as sequence of tense, noun denotations, 'Partee sentences' etc. To give just one example of how the proposal works, the problematic sentence 'Yesterday, John turned off the stove' will be represented using the composite tense operator 'It was the case yesterday that', which takes as input the tenseless content 'John turn off the stove' and shifts the evaluation time of the embedded content to a past time that belongs to the class of times picked out by the adverb 'yesterday', thus yielding the required truth conditions.

Brogaard's proposal is original and, if successful, would indeed offer an elegant treatment of the phenomena mentioned by King that is temporalist in nature. I have, however, a few worries about the proposal, the arguments for it and the general argumentative strategy of the book. The first worry concerns the syntactic evidence that is adduced in favour of the view that Brogaard presents, namely that tenses are circumstance-shifting sentential operators. To put it bluntly, I'm worried that there is not much positive evidence that is brought to support that view. The point is not that Brogaard doesn't appeal to empirical data — in fact, she considers the very same linguistic phenomena that King appeals to. But what she offers is a reinterpretation of the phenomena, rather than positive syntactic arguments for it. Is this, by itself, enough to counter King's allegations that temporalism cannot account for certain linguistic phenomena? Possibly so. But some true supporters of temporalism would want more than that: they would want positive, decisive arguments for the view that tenses are to be interpreted as circumstance-shifting sentential operators, rather than, say, quantifiers over temporal variables verbs come endowed with. The problem seems to me to be serious, given that Brogaard follows King in holding that

> 'the claim that tenses are operators that shift features of the index of evaluation is an empirical claim about natural language. It is a claim to the effect that in the best syntax and semantics for natural language, tenses will be treated syntactically and semantically as such operators.' (King 2003: 215, quoted in Brogaard 2012: 81).

As far as I can tell, no positive evidence that the view proposed is 'the best syntax for natural language' has been given.

A second, related worry concerns composite tense operators. As we have seen, Brogaard draws a distinction between basic tense operators and composite ones. A composite tense operator is the result of the syntactic combination between a basic tense operator and any of the temporal adverbial phrases mentioned above. Brogaard claims that the combination between such a phrase and a sentential operator is a novel, more complex sentential operator, and she is definitely right about that. But whether the composite tense operator is *sentential* depends entirely on the basic tense operators being sentential. However, if what I said above is right, no positive case that this is so has been made. One is thus tempted to ask: if there is no positive syntactic evidence for basic tense operators being sentential, what grounds do we have to claim that composite ones are sentential? What stops us from claiming, instead, that the combination of the basic tense operators with the various temporal adverbial phrases results in the creation of a complex adverbial phrase that might be interpreted not as a complex sentential operator, but, say, as a complex quantifier phrase? The point is not that composite tense operators won't solve the problems King and other have pressed against the temporalist; the point is that the account works only if it is presupposed that basic tense operators are sentential. But this, it seems to me, is what was at stake from the beginning; this is what temporalism requires and what its defenders were supposed to prove. Taking for granted that basic tense operators are sentential seems to put the cart before the horse, and prevents the temporalist to claim a clear victory over her competitors.

But perhaps Brogaard's forage into the syntax and semantics of tenses and temporal expressions is better seen as a defensive move against anti-temporalists such as King, rather than as purporting to offer positive arguments for the view. This opens up a more general issue related to the argumentative strategy Brogaard employs in her book. Usually, a viable operator treatment of tenses has been taken to give solid reasons for the postulation of time-neutral contents — that is, reasons for temporalism. This line of reasoning has been captured by what came to be known as 'the operator argument', the most prominent proponent and defender of which being Kaplan 1989. The success of the argument has been, of course, denied. (Brogaard tackles and defends this argument in Chapter 6.) But what I want to point out it that the strategy considered above can also be reversed. Instead of arguing from syntactic and semantic considerations about tenses and temporal expressions to temporalism, one could instead argue from other kinds of considerations to temporalism, and then propose an operator treatment of tenses and temporal expressions as fitting best with temporalism. We can find an example of such a reversed strategy in Recanati 2007: in his opinion, temporalism is best defended by considerations having to do with language learning and the regimentation of complex expressions in a simple language (2007: part 2), while the operator treatment of tenses is adopted as a consequence. Now, the question that arises in connection to Brogaard's book is the following: which is the direction of argumentation she favours? If my remarks above are on the right track, it is doubtful that the first direction will be successful. On the contrary, even if I am right about the above, the second direction is still available to her — and with significant results too. Of course, that puts additional dialectical weight on her other arguments for temporalism given in the book (that from belief possession and retention given in Chapter 2, that from disagreement given in Chapter 3, etc.). But assuming that those arguments are successful, the result that tenses should be treated as circumstance-shifting sentential operators can easily be achieved.[2]

[2] This presupposes that the operator treatment to tenses is the only one compatible with temporalism. There are, however, other views of tenses that are compatible with temporalism. The 'mixed view' hinted at in Recanati 2007 and developed in Zeman (ms.) is one example. So, following the second direction of argumentation described above doesn't strictly speaking lead us directly to a sentential operator treatment of tenses. I will set aside such complications in this note.

In sum, while I think there are weak points in Brogaard's arguments for an operator treatment of tenses and temporal expressions, the book as a whole contains many illuminating discussions of important issues pertaining to more than one area of philosophical inquiry. True, some of her positive proposals would benefit from a more developed treatment (such as, for example, the application of her proposal to the complicated linguistic phenomena dealt with in Chapter 5). This observation, however, is not a criticism of the book as such; it is, rather, an invitation to further develop the interesting points made in future work.

Dan Zeman
Universitat Pompeu Fabra
Department of Translation and Language Sciences
Roc Boronat 138, 6th floor, office 53610
08018 Barcelona
dan.zeman@upf.edu

## *References:*

Brogaard, B. 2012. *Transient Truths. An Essay in the Metaphysics of Propositions*. Oxford University Press.

Kaplan, D. 1989. Demonstratives. In *Themes from Kaplan*, ed. by J. Almog, J. Perry and H. Wettstein, 481-563. Oxford University Press.

King, J. 2003. Tense, Modality and Semantic Values. *Philosophical Perspectives* 17: 195-245.

King, J. 2007. *The Nature and Structure of Content*. Oxford University Press.

Marti, G. and Zeman, D. 2010. Review of Jeffrey King, 'The Nature and Structure of Content', Oxford University Press, 2007. *Mind* 119 (475): 814-19.

Recanati, F. 2007. *Perspectival Thought. A Plea for (Moderate) Relativism*. Oxford University Press.

Zeman, D. (ms.) Temporalism in an Extensional Framework. Retrievable at http://institutnicod.academia.edu/DanZeman/Drafts.

*Disputatio*'s Symposium on Berit Brogaard's *Transient Truths*
Oxford University Press, 2012

# On Two Arguments for Temporally Neutral Propositions

**Vasilis Tsompanidis**
Institut Jean Nicod and
Instituto de Investigaciones Filosóficas, UNAM

BIBLID [0873-626X (2013) 37; pp. 329-337]

In *Transient Truths*, Berit Brogaard 2012 offers a forcefully argued defense of what she calls 'temporalism': the view that many sentences express contents whose truth value can change over time. In recent years various arguments against temporalism have been posed by what seems to be the eternalist orthodoxy in analytic philosophy of language. I take Brogaard's book as the most complete and up-to-date reply to the eternalist attacks. Hence, I bypass this discussion here to critically examine two self-standing arguments she offers directly against the eternalist, and for temporally neutral contents. Sections 1 and 3 argue that at the moment the arguments are not entirely successful, while section 2 expresses strong doubts over Brogaard's choice of sentences such as 'John is a firefighter' to motivate and exemplify her view.

## 1 Disagreeing about occupations

The first argument against the eternalist starts from the known difficulty eternalists have with conversations that take place over extended periods of time. According to Brogaard, 'most of these conversations are not about specific times but about some other subject matter altogether', a subject she takes to be 'temporally neutral' (2012: 66). Her paradigm conversation is the following:

[FIRED_FIREFIGHTER]
A: … *John is a firefighter*
(Behind John's closed office door his superior is shouting 'You

Received: 03/10/2013 Accepted: 09/11/2013

are fired!')
B: I guess you are right. But *John is not a firefighter*. He was just fired.

Brogaard explains that B's claim 'you are right' sounds odd, but the eternalist translation of the conversation (1' below) is perfectly fine. Hence the eternalist translation is mistaken.

(1') A says that John is a firefighter at $t_1$, and B that he is not a firefighter at $t_2$.

Examples like [FIRED_FIREFIGHTER] seem an obvious problem for at least those eternalists that treat all present-tense verbs as referring strictly to the time instant when the sentence is uttered. The two conversants indeed seem to be talking about a more general subject matter. What could that be?

Well, it might be that they are talking about John's being a firefighter *at least up to* and *including* the time of *their entire conversation*. So when A utters

(1) John is a firefighter,

she might mean something like

(1*) John was a firefighter for some time before we started speaking, he is a firefighter now, and he will be a firefighter until at least we stop speaking.

If (A) means (1*) with her statement, B's retort 'I guess you are right' is actually wrong, since the last part of (1*) is not satisfied[1]. But now an eternalist can use (1*) to resist Brogaard's conclusion that the two conversants are sharing 'temporally neutral' information. (1*) is about a specific time interval, just a more complex one than it first

[1] There are of course real and difficult issues with specifying the exact meaning of phrases like 'you are right'. I will just note here the problematic use of a present-tensed form of the verb 'to be' in the phrase 'you *are* right': an insistent eternalist could reply that this is the exact issue the debate is trying to settle.

appears.

Note that (1*) is not the claim that 'B can freely choose which time his assertions refer to' that Brogaard correctly criticizes in (2012: 68-69). The choice of reference time is not free, as it is always restricted by at least the term 'now'. Nor does (1*) contain the problematic claim that John is a firefighter *at every time* over an extended time interval. As Brogaard notes, this is a non-starter but can be avoided by taking the interval as basic, and not every time instant in it. And neither is (1*) the claim that the referred interval stretches *unrestrictedly* into the future, a claim rightly attacked by Brogaard in page 71, since the time of (1*) is only supposed to stretch up until the point the conversations ends[2]. From the eternalist positions Brogaard mentions in her book, (1*) only comes close to the Salmon/Fitch intervalist positions, but it has some obvious restrictions on the referred interval that these positions do not seem to have, at least in the way Brogaard presents them.

## 2 On simple present tense sentences

There is one obvious drawback with (1*): it cannot be used as the final word on the semantics of every simple present-tense sentence. For example, it is not the intuitive analysis of the meaning of sentence 'John swims' truthfully uttered at a time when John is not actually swimming. In contrast, Brogaard's ultimate proposal in the last part of her book seems to offer a uniform treatment of sentences with verbs in the (grammatical) simple present tense. She claims, for example, that 'tensed sentences without time adverbials, when uttered at a particular time, do not make reference to the time of speech' (2012: 148), and, later, that 'temporal propositions are [...] the (natural) contents of simple present-tensed sentences without time adverbials' (2012: 155).

[2] There might of course exist *presuppositions* that the time referred can stretch more into the future — I return to this point in section 2. Incidentally, (1*) can be used for other sentences that Brogaard poses as problems for the eternalist, such as 'Mary loves me' (2012:42); [HAIR_SPLITTING] (72) and [FIRED_AGAIN] (70).

I will try to claim in this section that this issue might cut both ways. If it turns out that there is nothing simple about present tense sentences, and we have in our hands a case of widespread semantic polysemy or ambiguity that still needs to be investigated thoroughly, Brogaard's thesis can be denied. The eternalist could now reply that many of the intuitions and puzzles driving her case are there *exactly because* verbs in the present tense are ambiguous. Let me explain.

## 2.1 Occupations

I start with paradigm sentence (1) 'John is a firefighter'. It seems to me that there is not *one* reading of the sentence that conversants or philosophers of language can agree on, and that might form a stable basis of the agreement and disagreement intuitions Brogaard is after. A speaker can use (1) to talk about a person's current job (1a below), a person's past studies and intended occupation (1b), or even a person's 'call in life', what she and others define her as (1c):

(1a) John is a firefighter. He just now signed his contract with Firefighters United to fight fires.
(1b) John is a firefighter. He completed firefighter school in August and is now applying for jobs.
(1c) After 30 years with Firefighters United, John was fired last week and now works as a consultant. Despite that, he is, and always will be, a firefighter.

Each sentence above comes with its own temporal requirements, exactly because the properties we assign to John (current job, occupation, 'call-in-life') have different temporal lifespans. Accordingly, intuitions about 'believes that' sentences or possible conversational disagreements differ considerably. When we speak of current jobs, for example, we are very sensitive to the time of speech, while when we speak about an adult person's call-in-life we do not particularly care about it.

## 2.2 Being *x*

The issue is not specific to sentence (1), but can be extended to every use of the verb 'to be'. This is quite evident when we move to a different language such as Spanish, which has two different copula verbs: 'ser' and 'estar'.

The verb 'ser' is used to assign stable properties to an entity or class of entities, as in (the Spanish translations of) (2a) - (2c).

(2a) Angela is German.
(2b) *Kripke's piece is interesting.* (2012: 47)[3]
(2c) *My phone number is 283-1759.* (2012: 59)

These uses claim property stability over time- hence one could posit that they carry a semantic component, or pragmatic expectation, that the assigned property will be had by the entity for a long time in the future, or at least until very fundamental changes in the make-up of the entity take place. With 'ser', the eternalist (1*) reading I offered seems to me entirely appropriate.

In contrast, the verb 'estar' is used to assign temporary properties to an entity, such as location (3a), mood (3b), and current health (3c). It is also used for the present progressive as in English (3d).

(3a) *MARY: I am in Boston.* (2012: 46)
(3b) Vasilis is happy.
(3c) *John is pale.* (2012: 50)
(3d) John is swimming.

Here the standard eternalist translation tying the property attribution to the time of speech is closer to the meaning of the verb, and we might not even care about the property being part of the entity until the conversation ends. In these cases, however, eternalist meaning analyses such as (1') from section 1 are fine, and disagreement intuitions do not get off the ground. For example, the sentence

[3] Examples in italics are Brogaard's own examples to defend various claims for temporalism. For reasons of length, I do not get into how each example affects each claim, but a careful reader can do it if she so wishes.

utterer can insist that she *was* right at the time of speech, no matter what has happened by the time her conversant actually responds.

## 2.3 The present tense

Besides the idiosyncrasies of the verb 'to be', similar issues can be posed concerning the ambiguity (or polysemy) of the present tense construction of the English language.

First, the present tense has what have been called 'habitual' uses (4-6).

(4) Brit writes books.
(5) *Mary loves me*. (2012: 42)
(6) George is (works as) a summer tour operator.

Each of these sentences has at least one reading according to which the subject might not be $\varphi$-ing at exactly the time of the sentence's utterance, but has $\varphi$-ed with some regularity in the past, and is expected to carry on $\varphi$-ing in the future.

There are of course other uses of the present tense that might be taken to clearly, and only, indicate $\varphi$-ing at the time of utterance. And, at least in English, there are seemingly future-oriented uses of the present tense (7), and 'stable property assignation' uses (2a-c above).

(7) *I'm giving a talk in Alaska.* (2012: 154)

There might still exist more ambiguity or polysemy in the meaning of all tenses, to be revealed by further research in linguistics. My point here is simply that, given all this polysemy, it is too fast to claim that the (natural) contents of all the sentences I presented are temporal propositions. It might instead be that polysemy is exactly why the debate between eternalists and temporalists is puzzling, why we agree or disagree in specific examples, and why the eternalist 'translations' sometimes fail and other times sound incredibly obvious.[4]

[4] Note that Brogaard explicitly treats the term 'simple present tense' in her ultimate claims as denoting the 'grammatical present tense', not just one specific

Admittedly, a strict instant-based eternalist fares badly with the uncovered ambiguities; but even she could reply that her analysis is reserved for only one kind of present-tense uses. And eternalists in general should be able to survive Brogaard's criticisms by positing different temporal intervals that a sentence can refer to, depending on the exact verb in the sentence, and the present-tense use it captures.

## 3 Perception passes-on temporally neutral content

I now turn to what seems to me the most forceful stand-alone argument against the eternalist that Brogaard offers in her essay[5]. It starts from the quite plausible premise that 'the phenomenology of perceptual experience determines the content of mental states' (2012: 177). But the phenomenology of perceptual experience does not seem to discriminate among different times (2012: 176). Hence, the content of some mental states cannot contain a specific time among its constituents, and is thus a 'temporal proposition'. There is a lot to be said about this novel and important argument — here I just offer two reasons to be suspicious of the second premise.

### 3.1

Take a simple statement such as (8) below, intended to capture my perceptual belief, or *seeing-of*, a red car in front of me.

(8) There is a red car in front of me.

Most of our perceptual experiences are reported with such statements, so it does not seem far-fetched to conclude that perceptual experiences do not discriminate among different times.

But perceptual experience is far richer in the temporal domain than (8) suggests. First, the perceiver automatically includes the percept *in her present*, since she can also report it by (9) below.

use of it. Hence her claims cover all my section 2 examples.

[5] I focus here on the 'perceptual experience' part of her 8.4 argument only.

(9) There is now a red car in front of me.

In contrast to (8), (9) mentions a specific time, the one that 'now' refers to. One could argue just from this that some information about the specific time has to be part of the content of my perception, since (9) explicitly mentions it.

A perceiver always seems to know a lot about the temporal position of her percepts. I know, for example, that what I see is after my birth, concurrent with other percepts from the same or different modality, before or after others, and if a long time has passed since I registered it. That this happens ubiquitously in perception is shown by the fact that the information is easily recoverable — but it would not be if we just stored a 'temporal proposition' content such as (8) without any accompanying temporal data. Even if sometimes temporal phenomenology is very poor, by the time a perceptual belief is formed, it is put into a precise temporal position in the network of other past perceptions, current perceptions and desires, and future expectations.[6] The point is that, despite (8) not mentioning all this information, and it seeming indeed not to discriminate among different times, some information about the specific time of my perception is passed on to our perceptual judgments by perception, and thus might also be part of the judgments' content.

## 3.2

A second reason to be suspicious of Brogaard's second premise is that perceptual beliefs often do not seem to explicitly represent *locations* or precise *demonstrative information* either. So one might be allowed to form an analogous claim to her second premise: that when I believe, say, in Santa Barbara, and then inside a similar room in Paris, what I would express with the statement

[6] It is important to note here, in response to Brogaard's similar argument in (2012: 58), that the brain does not need to store the time name, say *2:30pm*, to keep track of the times of our perceptions. It only needs to store *de re* information that is specifically about that time. But then, pace Brogaard, the brain is a complex time-tracker, and, I should add, an incredibly efficient one at that.

(10) It is really hot,

my perceptions do not discriminate among the two locations. Similarly for beliefs I would express with the statement 'this cat is pretty', when looking at extremely similar cats that I nevertheless know are different. But the fact that the expression of the perceptual belief does not seem to include some obvious differences in information between the two perceptual instances does not mean that the information is not part of the content of my perception or perceptual belief. After all, this is why we have demonstratives and indexicals, to be able to refer to different situations with the same cognitive apparatus.

At this point Brogaard could retort that, despite the fact that location and demonstrative information is sometimes not explicitly represented in the sentences I use to express my perceptions, it *is* phenomenally available. This is shown by the fact that I do distinguish between a belief expressed with (10) that I am having in Santa Barbara, and one I am having in Paris. But this is exactly the move that I made on behalf of the eternalist in 3.1: that some information (for the eternalist, information about a specific time) can be part of perceptual content, without that meaning that it is able to be 'read off' from the sentences the perceiver uses to report on her perceptions. If this is true, the phenomenology of perceptual experience does discriminate among different times, and Brogaard's argument for temporalism does not go through.

Vasilis Tsompanidis
Institut Jean Nicod
UMR 8129
Pavillon Jardin
Ecole Normale Supérieure
29, rue d'Ulm
F-75005 Paris
tsompas@gmail.com

*Disputatio*'s Symposium on Berit Brogaard's *Transient Truths*
Oxford University Press, 2012

# Replies to Giuliano Torrengo, Dan Zeman and Vasilis Tsompanidis

**Berit Brogaard**
University of Missouri-St. Louis

BIBLID [0873-626X (2013) 37; pp. 339-352]

## 1 Reply to Giuliano Torrengo

I would like to start by thanking my commentators for their insightful comments, suggestions and objections. Their insights will no doubt help further discussion of temporalism and eternalism in the future and have already helped me make my own thoughts more precise. I will reply to their objections in the order that seemed most natural to me. Torrengo addresses the issue of whether temporalism has metaphysical implications, Zeman sets forth concerns of a methodological type and Tsompanidis raises objections to the book's main positive arguments. I will reply to my commentators in this order.

Torrengo addresses the interesting and very current question of whether the debate about temporalism versus eternalism has any bearing on the debate about presentism versus metaphysical eternalism. In the book I address this issue in a couple of places. One thing I say is that semantic eternalism seems inconsistent with presentism, a particular version of the A-theory. The argument is this. Presentism holds that only present things exist. But according to the standard version of semantic eternalism, all propositions include a timestamp (e.g., the sentence 'Mary is hungry' may express the proposition that Mary is hungry at 2:05 pm on October 1, 2013 CST). Most of these timestamps are past and future times. So, if presentism is true, then the vast majority of these propositions do not exist. The presentist could construe times as ersatz times (sets of propositions) (Brogaard 2013a). But on pain of circularity, this requires granting that there are temporal propositions (without a timestamp). So, presentism is at odds with semantic eternalism.

*Disputatio*, Vol. V, No. 37, November 2013
Received: 03/10/2013 Accepted: 09/11/2013

Torrengo replies that the argument doesn't work, because the view I argue for in the book is one that holds that there are some eternal propositions, for example, the propositions that there are wholly past objects and that I am giving a talk in L.A. on the 14th of November. Yet, Torrengo argues, 'it is the thesis that *some* eternal propositions exist that is at odds with presentism'.

This is a nice point. However, I disagree with Torrengo that presentism is at odds with the thesis that there are some eternal propositions. As he himself points out, it is the stronger view that there are no temporal propositions (i.e., semantic eternalism) that prevents the presentists from construing times as ersatz times. The weaker view defended in the book leaves us with all the resources (i.e., temporal propositions) needed to construe times as sets of propositions non-circularly. That said, Torrengo is perfectly right that if presentism is true, then the temporalist cannot accept all of the eternal propositions ordinary language appears to commit us to. In the book I argue (while bracketing metaphysical issues) that there are eternal propositions that make explicit reference to times, for instance, the proposition that I am giving a talk in L.A. on the 14th of November. If presentism is true, then that proposition does not currently exist. Presentists must, therefore, reject the existence of these kinds of propositions. (They can, of course, accept the existence of metaphysical propositions such as *there are wholly past objects*, as these types of propositions do not have times as constituents). The thought that sentences, such as 'I am giving a talk in L.A. on the 14th of November', do not express a proposition at all and therefore are false is not entirely unmotivated. It could be argued that while an utterance of the sentence 'I am giving a talk in L.A. on the 14th of November' may seem true, this kind of speech is, in fact, idiomatic much like 'the sun is rising'. Idiomatic speech is literally false (or untrue) but conveys something true.

Torrengo also raises an objection to my argument in Chapter 7 that if metaphysical eternalists adopt the quantifier account of the tenses (that is, the semantic eternalist's common account of the tenses), then they will have difficulties making certain metaphysical claims. The argument is too long to repeat here but the gist of it runs as follows. The metaphysical eternalist wants to say that past and future objects exist simplicer. Consider:

(1) Socrates exists.

Socrates existed in the past but does not presently exist. So, the metaphysical eternalist holds that (1) is true on one reading but false on another. Now combine metaphysical eternalism with the quantifier account of the tenses. On the quantifier account, all propositions are indexed to a time. So, where t* is the time of speech, (1) is equivalent to the proposition expressed by (2):

(2) Socrates exists at t*.

But here is the problem. If (2) specifies a false proposition expressed by (1), then what is the nature of the true proposition expressed by (1), according to the metaphysical eternalist?

Torrengo makes numerous very good points with respect to this argument. I will respond to what I take to be the main ones here (though in a different order). In response to my argument above, Torrengo argues that 'once we accept the distinction between a temporally restricted and a temporally unrestricted reading of quantification (and something analogous for predication), the worry is spurious'. However, this misses the point of the argument. The argument is that if the metaphysical eternalist accepts a quantificational account of the tenses, then she cannot account for the unrestricted reading of (1). (1) can, of course, be read as follows (as Torrengo suggests):

(3) $\exists x(x = \text{Socrates})$, where the domain of values is temporally unrestricted.

(4) $\exists x(x = \text{Socrates})$, where the domain of values is restricted to the present.

According to the metaphysical eternalist, (3) then is true and (4) false. However, this proposal is compatible with a version of temporalism that utilizes quantifier restriction. My own proposal was similar. On the view I prefer, (1) has a reading that determines a function from worlds to extensions and another reading that determines a function from world-time pairs to extensions. The first reading is the "unrestricted" reading, whereas the second is the "restricted" reading.

Notice, however, that neither of these proposals utilizes a quantifier account of the tenses. In fact, they are inconsistent with the standard version of semantic eternalism, which requires that all propositions are indexed to a time. And that was just the point of the argument in Chapter 7, which was not an argument against metaphysical eternalism but one in favor of temporalism (on the assumption that metaphysical eternalism is true).

A second worry that Torrengo raises is that the presentist cannot coherently claim that (1) is false, on an unrestricted reading. The reason for this, he says, is that I hold that an eternal proposition such as *Socrates exists* is 'evaluable as true or false *simpliciter* only in context in which either Socrates is a instantaneous object, or Socrates always (or never) exists (150). However, this is not my view. What I said was:

> 'I think that one *could* use 'John has a straight shape' to mean the eternal proposition that John has a straight shape. But such an eternal claim is truth-evaluable at a world *w* only if (i) John is an instantaneous object at *w*, (ii) John always has a straight shape at *w*, (iii) John never has a straight shape at *w*, or (iv) Lewis is right that the *eternal* proposition *John has a straight shape* is true at *w* iff John has a temporal part that has a straight shape' (Brogaard 2012: 150).

I made this remark in the context of discussing Lewis's problem of temporary intrinsics. The reason 'John has a straight shape' cannot be evaluated except under these conditions is that if John sometimes has a straight shape and sometimes has a bent shape, then relative to the world as a whole the proposition is neither true nor false (or both true and false). The same point does not apply to the proposition that Socrates exists (as existence does not come and go).

A third concern that Torrengo raises also concerns the unrestricted reading of sentences like (1). He argues that my view implies that the unrestricted readings of sentences are never true for the presentist, not even when the entity in question is present. This has the consequence, he says, that 'the presentist and the eternalist necessarily disagree on what *presently* exist', which seems odd.

I agree with Torrengo that that would be odd. However, I don't think I am committed to this view. Consider:

(5) Obama exists.

If presentism is true, then (5) is true when read restrictedly and unrestrictedly. On the restricted reading, 'Obama' determines a function from world-time pairs to extensions. Since the extension is non-empty, (5) is true on this reading. On the unrestricted reading, 'Obama' determines a function from worlds to extensions. Since this extension is also non-empty, (5) is true on this reading. Torrengo thinks I cannot say this, because I argue that on the unrestricted reading, (5) entails that it will be the case that Obama exists. However, even if we bracket Obama's future existence, there is no problem here, because this kind of tensed sentence is innocuous. It is the result of affixing a tense operator to a sentence given an unrestricted reading. But when tense operators are affixed to an operand sentence that expresses an eternal proposition, the tense operators will be redundant (150). So, the presentist can agree with the metaphysical eternalist that (5) is true on both its restricted and unrestricted reading.

Torrengo is right that if the presentist holds that Obama is not fully present but is unfolding in time, then it would seem that she should reject (5). After all, if only some of Obama's parts exist, how could (5) be literally true? I think this is a genuine puzzle but not one that is specifically about the unrestricted reading of (5). It appears to be equally problematic on the restricted (ordinary) reading of (5). However, the puzzle is not a consequence of accepting presentism or temporalism. Anyone who holds that ordinary material objects are four-dimensional spacetime worms needs a way to talk about the properties the present parts instantiate. This is a familiar issue from the metaphysical literature (see e.g., Sider 2001). It is true that Obama is speaking even if it's only his present part that is speaking. Yet how can this be if he is extended four-dimensionally? One standard reply is that proper names ordinarily refer only to stages of objects. Whether this is the best reply to the worry is not something I can address here. But let me point out that most three-dimensionalists who take ordinary material objects to endure are faced with a version of this problem. It is commonly agreed upon that events perdure: they have temporal parts located at different times. Yet even if a soccer match takes a considerable amount of time, it can nonetheless still be true to say that you are currently watching one. So, the problem of how to correctly predicate properties of four-dimensional entities may arise regardless of one's particular view of how ordinary material objects persist through time.

## 2 Reply to Dan Zeman

Though temporalism is not formulated as a view about how to treat the tenses in English, I argue in the book that on the most natural understanding of temporalism, the debate between temporalism and eternalism is not orthogonal to the debate about how to treat the tenses. Standard versions of eternalism require that the time of speech is a constituent of all propositions. As the time of speech is variable, sentences that express eternal propositions must have a hidden variable in the sentence structure that takes times of speech as its values. This type of sentence structure follows as a natural consequence of a treatment of the tenses as quantifiers. Where 't*' is a variable that takes times of speech as its values, 'John is a firefighter' is of the form 'John is firefighter at t*', 'John was a firefighter' is of the form 'there is a time t such that t is earlier than t*, and John is a firefighter at t', and 'John will be a firefighter' is of the form 'there is a time t such that t is later than t*, and John is a firefighter at t'.

Temporalism, by contrast, must treat the tenses as sentential operators, at least given standard semantics. It may be thought that it is possible to combine temporalism with a quantificational account of the tenses. For example, it may be thought that 'John was a firefighter' could be treated as having the following underlying form:

(6) $\exists t(t < t_n$ & John is a firefighter at t),

where $t_n$ is an unarticulated constituent that takes different values across time. If (6) expresses a proposition with an unbound variable, then that proposition will have different truth-values at different times. The problem with this view is that a content that contains an unbound variable isn't a complete proposition, given standard semantics. In standard semantics sentences, relative to context, express complete propositions that to not require further satisfaction by context. So, unless we adopt some special semantics, (6) expresses an eternal proposition, viz. the proposition that results from substituting the time of speech for $t_n$. It thus seems that within a fairly standard semantic framework, temporalism is committed to a treatment of the tenses as sentential operators, whereas eternalism is committed to a treatment of the tenses as quantifiers over times or some similar view (e.g., a treatment of the tenses as quantifiers over events or as discourse variables).

As Zeman points out in his commentary, a treatment of the tenses as quantifiers is a minority view among linguists (philosophers have been far more more sympathetic to it). In the book I reply to a number of the arguments that linguists and philosophers, like Jeff King 2003, have offered against the operator account. Zeman raises several novel, worries about the temporalist's suggestion that we treat the tenses as operators. He grants that it may be the case that one can come up with an operator account of the tenses that can accommodate most, if not all, of the phenomena that normally are cited in support of the quantificational account. However, he believes that true supporters of temporalism might want 'positive, decisive arguments for the view that tenses are to be interpreted as circumstance-shifting sentential operators, rather than, say, quantifiers over temporal variables verbs come endowed with'.

Zeman then suggests that in the absence of any such positive arguments that show that the tenses are best treated as operators rather than as, say, quantifiers, a different argumentative strategy may be more efficient. The different argumentative strategy Zeman proposes is to provide compelling, independent support for temporalism and then show that temporalism requires a treatment of the tenses as operators.

I agree with Zeman that there are very few empirical data concerning the semantics of tense that cannot be accommodated by both operator accounts and quantificational theories of the tenses (as well as many other theories of the tenses). So, my reply is not going to be to become up with a range of new empirical data supporting the operator account. In fact, I completely agree that the second strategy is the only strategy that is going to work for the temporalist. However, I also took that to be the strategy of the book. I think temporalism offers a better account than eternalism of belief retention (Chapter 2), agreement and disagreement across time (Chapter 3) and the phenomenology of perceptual experience (Chapter 8). Here I will provide a quick overview of the argument from the phenomenology of perceptual experience with some emphasis on the argumentative strategy that Zeman and I both agree is the best strategy for the temporalist. Tsompanidis offers some independent objections to this argument. I will revisit those below.

My argument for temporalism based on the phenomenology of perceptual experience made a couple of assumptions that some may find controversial, viz. the assumptions that (i) visual experiences have propositions as their content, and (ii) the content of visual experience supervenes on the phenomenology. However, a version of the argument can be made without these assumptions in place. The original argument went as follows. Two subjects could, in principle, have phenomenally identical experiences at different times. If the phenomenology of visual experience determines the content, then two subjects could, in principle, have experiences with the same content at different times. It follows that times are not constituents of the contents of experiences. Given the assumption that the content of experience is a proposition, there are propositions that do not have times among their constituents. So, there are temporal propositions.

While the argument, as formulated, will only be compelling to those who already accept the two assumptions about the content and phenomenology of experience, a version of the argument establishes the same conclusion without these two assumptions in place. Qualia theorists reject the view that the phenomenology of experience supervenes on the content but not the converse (Block 2010). The main, current view that will reject the assumptions I originally made is naive realism. However, there is a different argument for the same conclusion that assumes naive realism. Naive realists normally hold that the external object that triggered the experience and its visually perceivable property instances fully constitute the phenomenal character of experience (Martin 2002; Campbell 2002; Brewer 2007; Kennedy 2009). As times (such as 2 pm today) are not normally visually perceivable properties, there could be two phenomenally identical experiences that are temporally distinct. For example, an experience of one and the same red tomato on two different occasions. Moreover, if the subjects of those experiences were to describe as sincerely as possible what their experiences convey, their descriptions would not make any reference to times. If they did, there would be more accurate descriptions not making any reference to times. The descriptions convey propositions, despite making no reference to times. Hence, there are temporal propositions.

Returning to Zeman's point, I agree with him that the temporalist needs to rest her case on considerations that are independent of arguments about the correct semantics of tense. I also think considerations like the one above about visual experience make temporalism seem more appealing than eternalism. But temporalism requires that one treats the tenses as operators. So, if the operator account and the quantifier account of tense can both accommodate the empirical data about tense, which appears to be the case, then we should favor the operator account.

## 3 Reply to Vasilis Tsompanidis

Tsompanidis raises some interesting objections to two of the book's main positive arguments for temporalism. His first point of contention is with my argument that temporalism is better suited as a semantics of agreement and disagreement. As Tsompanidis points out, the argument rests on cases of the following kind:

> [FIRED FIREFIGHTER]
> A: … *John is a firefighter*
> (Behind John's closed office door his superior is shouting 'You are fired!')
> B: I guess you are right. But John is not a firefighter. He was just fired.

B's claim 'you are right' sounds odd, but the eternalist translation of the conversation is perfectly fine. So, the eternalist translation is mistaken.

Tsompanidis raises several objections to this type of argument. I will briefly review the two main ones here and then offer a reply to the first. The first point is that the eternalist could turn to interval semantics to account for agreement and disagreement. For example, 'John is a firefighter' might mean 'John is a firefighter *at least up to* and *including* the time of *the entire conversation*'. This type of account may be able to explain what is wrong with dialogues like the one presented in [FIRED FIREFIGHTER]. The second point is that ‘to be’ in the present tense is polysemous and hence may yield different contents in different linguistic contexts. Tsompanidis notes that it may be that polysemy is ‘why the debate between eternalists and temporalists is puzzling, why we agree or disagree in specific examples, and why the eternalist ‘translations’ sometimes fail and other times sound incredibly obvious’.

Tsompanidis makes many very good points, and unfortunately I cannot reply to all of them here. However, let me consider the first point that eternalism could take the present tense to refer to intervals. As Tsompanidis notes, I do consider this kind of reply at length in the book but let me address the specific account he proposes. One major problem for defenders of this type of proposal is to give precise truth-conditions for sentences, given that conversations do not have clear boundaries. A further, related, problem is that the time of the entire conversation cannot always serve as a reference time. Consider the following sentences:

(7) (a) Mary is falling down from the tree.
(b) Afghanistan is at war.
(c) I am alive.

If 7(a) is uttered during an extended conversation that may continue for hours while Mary is taken to the hospital, the relevant time interval cannot be one that includes the entire conversation. In this case, it may be suggested that the time interval is determined by the duration of the event. However, this suggestion cannot be right. I might utter 7(a) because I believe that Mary is falling down from the tree, even though she is not. In that case, there is no event to determine the relevant time interval. While there are many other proposals that could be considered, the sentences in 7(a)-(c) suggest that it will be difficult to give a systematic account of the time intervals that the present tense is supposed to make reference to. Though I agree with Tsompanidis that there are very many points that need to be settled about how language makes reference to time, I think that the problems the eternalist encounters with respect to agreement and disagreement give us a strong reason to prefer temporalism to eternalism.

Tsompanidis also addresses the argument for temporalism based on the phenomenology of visual experience. His main concerns lie with the second premise, viz. the premise that the phenomenology of perceptual experience does not seem to discriminate among different times. The first of Tsompanidis's reasons for not blindly accepting this premise is anchored in how we report on our visual experiences. Upon seeing a red car I might say:

(8) There is a red car in front of me.

However, as Tsompanidis correctly points out, it would hardly be inaccurate to use (9) as a report of my perceptual experience instead of (8):

(9) There is a red car in front of me right now.

'Now' makes reference to the time of speech, which we can stipulate is also the time of perception. But if (8) and (9) are equally adequate reports of the content of my visual experience, and (9) makes reference to a time, then it might be argued that there are times in the content of perception.

I think that this is a valid point about how we report experiences. However, I think there is reason to believe that (8) is a more adequate way to report the phenomenology of my experience than (9). The main reason to doubt that the phenomenology (and content) of visual experience involves times is that we can have phenomenally indistinguishable experiences across time. If I were to tie you up in front of a blue wall, ensuring that you could not move any part of your body except your eyes, the phenomenology of your visual experience would not change. Although time would pass, you would continue to have an experience with the same phenomenology. So, if the phenomenology of experience determines its content, you continue to have an experience with the same content. But if the phenomenology and content of your experience involved times, then you would not continue to have an experience with the same phenomenology and content. So, it seems that the phenomenology and content of your experience do not involve times. These types of considerations suggest that (8) is a more accurate report of the phenomenology of the envisaged experience than (9).

I agree with Tsompanidis, of course, that we can perceive many temporal aspects of reality. For example, even if your visual experience doesn't change in the envisaged scenario, it is very likely that you perceive time as passing. We also perceive certain events as occurring before others, and we perceive everything as occurring in the present moment. I wholeheartedly embrace those facts about perceptual experience. My argument only rests on the idea that the phenomenology of visual experience *need not* change, even though time is passing. This suffices to establish the second premise of the argument.

Tsompanidis's second objection to the argument from visual experience is that experience and perceptual beliefs 'often do not seem to explicitly represent *locations* or precise *demonstrative information* either'. But at first glance, at least, it would seem odd to deny that this type of information is conveyed by our perceptual experiences and our perceptual beliefs. For example, if I experience a red object, the content of my experience seems accurately captured using a report like 'that is red'. The demonstrative 'that' refers directly to a concrete particular, which suggests that a concrete particular is a constituent of the content of my visual experience. Indeed, many theorists would argue that particulars (objects and their visually perceptible property instances) exhaust the phenomenology of visual experience. But the point I made about times seems equally applicable to the case of material objects: If the object I happen to be looking at had been replaced by an indistinguishable object, it would introspectively seem as if I had an experience with the very same phenomenology.

The point is well taken. However, unlike many others I doubt that external, material objects are constituents of the contents of our visual experiences. So, the feeling that visual experience makes direct reference to an external object is not indicative of any actual direct reference. In fact, I provide a generalized version of the book's argument elsewhere (Brogaard 2013b). This raises the question of how beliefs sometimes come to refer directly to times, locations and material objects. For example, if I experience something red, then how do I come to believe directly of a particular object that it is red? The quick answer is that I take perceptual contents to involve self-locating constituents (i.e., constituents that have extensions only relative to centered worlds). Where /that/ is a self-locating constituent that refers to o (the object demonstrated), my visual experience that /that/ is red can be a cause and justifier of my belief of o that it is red.

Although I reject the view that visual experience makes direct reference to external objects, I want to emphasize that one could deny that the phenomenology of visual experience involves times yet nonetheless think that it is exhausted by material objects and their visually perceivable property instances. In my reply to Zeman I presented an argument for the view that the phenomenology of visual experience is not time-involving, on the assumption that naive realism is true: The naive realist holds that the phenomenology of visual experience is exhausted by the external object and the visually perceivable property instances of that object. Times are not normally among the perceivable features of a visual scene. So, the phenomenology of visual experience does not normally involve times. It is thus open to argue that the phenomenology of visual experience involves material bodies (and even locations) and yet deny that it involves times.

Berit Brogaard
University of Missouri-St. Louis
Department of Philosophy
599 Lucas
1 University Boulevard
St. Louis, MO 63121-4499
314-516-5631
brogaardb@umsl.edu

## *References*

Block, N. 2010. Attention and mental paint. *Philosophical Issues*, 20 (1): 23-63.

Brewer, B. 2007. Perception and its objects. *Philosophical Studies,* 132: 1, 87–97.

Brogaard, B. 2012. *Transient Truths: An Essay in the Metaphysics of Propositions*. New York: Oxford University Press.

Brogaard, B. 2013a. Presentism, Primitivism and Cross-Temporal Relations: Lessons from Holistic Ersatzism and Dynamic Semantics. In *New Papers on the Present: Focus on Presentism*, ed. by Roberto Ciuni, Kristie Miller and Giuliano Torrengo, 253-280. Philosophia Verlag.

Brogaard, B. 2013b. It's not what it seems. A semantic account of 'seems' and seemings. *Inquiry* 56/2-3: 210-239.

Campbell, J. 2002a. *Reference and Consciousness*, Oxford: Oxford University Press.

Fish, W. 2009. *Perception, Hallucination, and Illusion*. New York: Oxford University Press.

Kennedy, M. 2009. Heirs of nothing: the Implications of transparency. *Philosophy and Phenomenological Research* 79, 3: 574-604.

King, J. 2003. Tense, Modality, and Semantic Values. In *Philosophical Perspectives* 17: *Language and Philosophical Linguistics*, ed. by John Hawthorne and Dean Zimmerman, 195-246

Martin, M. G. F. 2002. The Transparency of Experience. *Mind and Language* 4: 376- 425.

Sider, T. 2001. *Four-Dimensionalism: an Ontology of Persistence and Time*. Oxford University Press.

# Book reviews

**Epistemological Disjunctivism**, by Duncan Pritchard. Oxford : Oxford University Press, 2012, 206 pages.

BIBLID [0873-626X (2013) 37; pp. 353-360]

In *Epistemological Disjunctivism*, Duncan Pritchard aims to present a McDowell-inspired version of epistemological disjunctivism and to defend it against three main criticisms. The central claim of Pritchard's view, which he restricts to perceptual knowledge, is that such knowledge is 'paradigmatically constituted by a true belief whose epistemic support is both factive [...] *and* reflectively accessible to the agent' (2-3). Pritchard does not offer positive arguments for this view but since, as he claims, it has the potential to be the 'holy grail' of epistemology, this should prove to be enough motivation for epistemologists to take the view seriously. So, although the goals of *Epistemological Disjunctivism* (*ED*) might seem modest, the project is worth undertaking given that many, if not most, epistemologists think that the view is clearly false. Importantly, Pritchard skillfully succeeds, in a rather short book, in placing epistemological disjunctivism (ED) as a position worth considering.

The book is divided into an introduction and three parts that consist of three substantial essays that are partly based on previously published work by Pritchard. The book starts properly in Part One, but it begins with a brief introduction where the above initial statement of the view is provided and where the two main theoretical benefits that render the view attractive are introduced. One benefit concerns the debate between epistemic internalism and externalism. The view seems to combine aspects found in the internalist and externalist approaches in such a way that, Pritchard suggests, it offers a legitimate third option to them and deals with both camps' main challenges. The other benefit concerns radical scepticism. Given that the view enables us to have reflective access to reasons that entail facts about the world, radical scepticism can lose its bite and Part Three is dedicated to show us how this can be done.

In Part One, on the other hand, Pritchard occupies himself partly with the first benefit, as well as setting out in more detail what the view amounts to, introducing three *prima facie* problems facing ED and considering two of them. Here Pritchard articulates the *Core Thesis* of ED as follows:

> 'In paradigmatic cases of perceptual knowledge, S has perceptual knowledge that *p* in virtue of being in possession of rational support R for her belief that *p* which is both *factive* (i.e., R's obtaining entails *p*) and *reflectively accessible* to S' (13)

Paradigm cases of perceptual knowledge are cases where S sees that *p* and believes that *p* on this basis and S has no undefeated psychological defeaters concerning *p*. So, in such paradigm cases, the rational basis for the belief that *p* has two interesting features.

First, the rational basis is factive. So beliefs about the external world can be based on factive reasons: reasons that imply that the beliefs are true. So a central thesis of ED is that you can have factive support for your beliefs about the external world.

Second, this epistemic support can be reflectively accessible. One can have reflective access to the fact that one sees that *p*, where this reflective access is understood in terms of what can be known through reflection alone (i.e., through a priori reasoning and introspection).

The disjunctivist aspect of the view is brought to light by comparing the sort of epistemic support one would have in paradigmatic cases and in bad cases where S does not see that *p* but S's experiences are introspectively indistinguishable from the ones in the paradigmatic case. The orthodox view suggests the subject has the same degree of reflectively accessible rational support in both cases, whereas ED holds that this support is radically different in kind: the subject possesses factive support in the paradigmatic case, but she lacks this sort of support in the bad case.

So, given the above reflective accessibility requirement, ED is committed to *accessibilism*: S's epistemic support is constituted solely by facts that S can know by reflection alone. Moreover, given the above disjunctivist approach, ED rejects the *new evil genius thesis* that internalists normally accept: S's epistemic support is constituted solely by properties that S shares in common with her envatted physical duplicate. So ED is not a paradigmatic internalist view (i.e., a view that maintains the strong supervenience of epistemic support on the internal). Nevertheless, Pritchard believes it has the means to capture our intuitions about epistemic responsibility since the epistemic support is within one's reflective ken and importantly, given the factivity of reasons in the paradigmatic cases, it pre-empts externalist worries concerning the truth-connection.

Pritchard also tries to provide some pre-theoretical motivation for the view in Part One. Worryingly, it is rather feeble: ED accommodates some of our ordinary way of talking about these matters. But it is not Pritchard's intention in *ED* to convince us that ED *is* true, but that it is not *plainly* wrong. So, he introduces what he considers to be the three key sources of dissatisfaction with ED, which are the main concern of this review.

The first is a sort of McKinsey-style problem that he calls the *access problem*:

P1. I can know via reflection alone that my reason for believing the specific empirical proposition *p* is the factive reason R.
P2. I can know via reflection alone that R entails *p*.
C. I can know via reflection alone that *p*.

But (C) seems false, since *p* is a contingent proposition about the external world. So, given (P2) seems true, it seems that (P1) is false. But, ED seems to entail (P1).

Nevertheless, as Pritchard shows, once one realizes that ED is not committed to the claim that one can have knowledge of specific empirical propositions from reflection *alone*, given that R is an *empirical* reason that one *sees that p*, the problem vanishes. In fact, the alleged problem is so easily solved that it is difficult to imagine that it could be (even partly) responsible for anyone's reluctance to accept ED.

A more serious source of dissatisfaction is the *basis problem*. Roughly, the problem is that the following three claims seem inconsistent:

i. If one sees that *p*, one knows that *p*.
ii. Seeing that *p* is the epistemic basis for knowing that *p*.
iii. One's epistemic basis for knowing that *p* cannot entail that one knows that *p*.

Given that ED is committed to (ii) and that (i) seems very plausible (seeing that *p* seems to be a way of knowing that *p*), it seems that the basis for knowing that *p* is knowing that *p*. But, given (iii), that cannot be right.

Pritchard wants to deny that seeing that *p* is a specific way of knowing that *p*. And he rejects the *entailment thesis* (i) by means of counter-examples. One such case exploits misleading defeaters. One sees a barn but is also told that one is in the land of fake barns. One does not know then that there is a barn before one, given the undefeated defeater. But suppose that later on one discovers that the testimony was false, would one retrospectively treat oneself as having *seen that* there was a barn? Pritchard thinks so and that might just be right. So, it seems that one can see that *p* without knowing that *p*.

Nevertheless, Pritchard holds that seeing that *p* necessarily puts one in a good position to gain knowledge that *p* even if one cannot exploit the opportunity. This, he claims, allows us to capture part of the insight that motivates (i): that seeing that *p* is both factive and *robustly epistemic*. So, Pritchard thinks he has resolved the basis problem.

But there seems to be a related issue lurking in the background. After all, Pritchard's story seems to go roughly like this: having reflective access to the fact that one sees that *p* is the rational support for which you believe that *p*, which in turn explains how you come to know that *p*. Now, it seems right to suggest that: if one knows by reflection that one sees that *p* (i.e., that one is in a factive state that has *p* as content), one knows that *p*. And, given how Pritchard seems to understands reflective access, it seems that is what he is suggesting. Say, if I can tell by introspection or reason that I see that there is a desk before me, I know that there is a desk before me.

But if this is so, the story we got hides a deficiency connected to another way of knowing. It is true that Pritchard is only interested in *perceptual* knowledge, but it seems that a complete story as to how we can perceptually know will need to say something about how we can know via introspection or reason. Frustratingly, we do not get that story.

The third source of concern Pritchard introduces in Part One is the *distinguishability problem*. It seems that the following two claims are in tension:

a. I have reflective access to factive reasons in the paradigmatic case.
b. The paradigmatic case is indistinguishable from the bad case.

This is because it seems that if one has reflective access to something in a case, one should be able to exploit it to distinguish this case from cases in which this thing is not present. So, it seems that if one claim is right, the other is wrong. And Pritchard dedicates the whole of Part Two to resolve this problem, since doing so requires motivating a distinction between favouring and discriminating epistemic support.

So, Part Two has two goals. First, to show that there is an independently motivated and plausible distinction between favouring and discriminating epistemic support. Second, to show that such distinction can be exploited to avoid the distinguishability problem. Let's consider the first goal.

The widespread *core relevant alternatives intuition*, that states that it is a necessary condition of knowing that *p* that one is able to rule out all relevant not-*p* alternatives, and the intuitive connection between perceptual knowledge and discrimination suggest an attractive *Relevant Alternatives Account of Perceptual Knowledge*, where to rule out an alternative is to be able to make the relevant perceptual discrimination:

> 'S has perceptual knowledge that *p* only if S can perceptually discriminate the target object at issue in *p* from the objects at issue in relevant (not-*p*) propositions, where a relevant alternative is an alternative that obtains in a near-by possible world.' (67)

But a problem of this account is that if one holds a plausible closure principle, perceptual knowledge becomes very hard to come by. In other words, given this account and closure, it seems one should be able to know things that intuitively one should not be able to know. For example, by knowing that one is looking at a zebra one should also be able to know that one is not looking at a disguised mule.

Pritchard, like most epistemologists, is not willing to give closure up (nor to become an sceptic). Instead, he wants to reject the view that to rule out an alternative is to possess the relevant discriminative ability, while retaining the spirit of the relevant alternatives account. To do this, we need to construe the evidence available in these cases broadly enough to include background information. This way one can have favouring epistemic support to think that what one sees is a zebra rather than a disguised mule, even though one cannot perceptually discriminate between those objects. So, we have two kinds of epistemic support: one provided by favouring evidence, the other by discriminatory capacities. And Pritchard thinks that it is this overlooked distinction that allows us to make sense, as closure suggests, that one can know that one is not looking at a disguised mule although one does not possess the relevant perceptual discriminatory capacities.

Unlike other theories, Pritchard's rightly does not hold that the background evidence is required to know perceptually, say, that one is looking at a zebra. If one does not become aware of the disguised-mule possibility, one can know simply by means of one's discriminatory powers. However, if one becomes aware of the error-possibility, one can know only if one can eliminate that alternative by means of favouring evidence.

So, given closure, if one knows that it is a zebra and one becomes aware of the disguised-mule possibility while making the competent deduction, one *must possess* the background evidence that allows one to know that it is not a disguised mule. Of course, one might not possess this evidence, but in that case, Pritchard suggests, one does not retain the knowledge that it is a zebra because one is unable to rationally dismiss the error-possibility that you know is incompatible with what you believe. So there is no violation of closure.

However, it is not clear that all such *salient* error-possibilities need to be rationally dismissed in order to know. Indeed, this seems to be an unwelcome result. Salient error-possibilities need not be regarded as relevant ones (especially since unmotivated challenges are usually conversationally illegitimate; 146) and it seems possible that one does not regard the error-possibilities as relevant even if one does not have the means to rationally dismiss them (cf. sceptical error-possibilities merely raised). But if this is so, closure does not hold in full generality and so the motivation for the distinction between favouring and discriminating epistemic support starts to wane.

But there might anyway be some such distinction that ED can exploit to its benefit. Given such distinction, one can reflectively distinguish between the paradigmatic and bad cases either by being in possession of favouring evidence or a discriminatory capacity. Overlooking this distinction is what makes (b) plausible, when thinking that knowing the difference implies knowing the difference discriminatorily. And although one cannot discriminate between the cases, in the paradigmatic case, one can anyway know that one is in such scenario as opposed to the bad one by being in possession of favouring evidence. So, if we understand (b) with the help of the above distinction, the tension vanishes, and so the final problem Pritchard considers, *if* we are willing to accept that one always has grounds for dismissing the error-possibility in the paradigmatic cases.

In Part Three, Pritchard moves towards the application of ED to the problem of radical scepticism. He is mainly concerned with the following argument:

BIV1. I don't know that I'm not a brain in a vat (BIV).
BIV2. If I know that I have two hands, I know that I'm not a BIV.
BIVC. I don't know that I have two hands.

(BIV1) seems plausible given the nature of the BIV-hypothesis and (BIV2) is motivated by a closure principle that seems plausible to Pritchard and many others. Of course, closure might not hold in full generality even with Pritchard's distinction between favouring and discriminatory epistemic support in place, but leaving that aside, we seem to have reason to worry about this argument.

ED is a form of neo-Mooreanism and so rejects (BIV1). The reason (BIV1) is false, Pritchard thinks, is that: given sceptical error-possibilities are not epistemically motivated, one can rationally dismiss the sceptical hypothesis, in paradigmatic cases, simply by means of one's available reflective access to factive rational support for the relevant belief.

But (BIV1) does not seem false and in fact the Moorean assertions (e.g., 'I know I'm not a BIV') do. So Pritchard also needs to give a story as to why this is so. Since he thinks that, typically, explicit (perceptual) knowledge claims represent oneself as possessing the relevant discriminating evidence rather than favouring evidence and one does not possess the relevant discriminating evidence due to the nature of the sceptical challenge, it seems conversationally inappropriate to make the true Moorean assertions.

So, ED's philosophical potential should be apparent and Pritchard's treatment of these and other issues is very rich (far more than what has been here portrayed). *ED* makes a concise but strong case to place ED as a 'live' option in the epistemological terrain. So *ED* beautifully succeeds in achieving its main goal and it should be read by anyone with an interest in epistemology.

Leandro De Brasi
Departamento de Filosofía
Facultad de Filosofía y Humanidades
Universidad Alberto Hurtado
Alameda 1869, piso 3
Santiago, Chile
ldebrasi@uahurtado.cl

## **The Transactional Interpretation of Quantum Mechanics**, by Ruth Kastner. Cambridge University Press, 2013, 224 pages.

BIBLID [0873-626X (2013) 37; pp. 361-367]

What does it mean to say that an event is physically possible? One might say, in the spirit of everyday usage, that an event is physically possible if its occurrence is consistent with the laws of physics. But which laws of physics, starting from when? Here things get tricky. In a Laplacian universe governed by the fully-deterministic laws of classical electrodynamics and mechanics, exactly one future is consistent with any complete and precise specification of initial conditions. Given an incomplete, imprecise specification — the best that human observers can achieve, even in principle — many futures may appear to be physically possible, but only one of them actually is. In this kind of universe, our intuitive notion that many different outcomes of an experiment, for example, are physically possible is just an ignorance-driven delusion.

Philosopher of physics Ruth Kastner's new book, *The Transactional Interpretation of Quantum Mechanics*, subtitled *The Reality of Possibility*, takes on this question of the meaning of physical possibility in the theoretical context defined by relativistic quantum field theory. Lest this seem daunting, be assured that this is a philosophical book; aside from Chapters 5 and 6 dealing with technical objections, it emphasizes ontological questions and employs the physics — all of which is clearly presented for the non-expert — as a way to raise them. The ontological questions that it raises, not just 'what is possibility?' but also 'what is actuality?' and 'what are space and time?' underlie not just physics but any theory of action. The *Transactional Interpretation of Quantum Mechanics* is, therefore, not just about the interpretation of quantum mechanics; it is about how 'nature makes its choice' in any circumstance, and how 'the actual arises from the potential' (205) as a result.

After briefly introducing quantum theory and its various interpretations in Chapter 1, Kastner addresses, in Chapter 2, the essential preliminary question of what is to count as an acceptable scientific explanation. The divide between realists and instrumentalists about quantum theory — or between 'ontic' and 'non-ontic' approaches to its physical interpretation — turns on this basic question. Kastner comes down explicitly on the side of Einstein, Bohm, Everett, Bell and other realists: she adopts as a 'maxim' that 'mathematical operations of a theory which are necessary to obtain correspondence of the theory with observation merit a specific (exact) ontological interpretation' (37). With this, she rejects the notion that quantum theory is just about what we can know, a position associated historically with Bohr and Heisenberg and advocated more recently by Peres, Bub, Fuchs, Spekkens and others. While Kastner might be criticized for treating Bohr as too straightforwardly Kantian — she characterizes him as 'pre-emptively denying that the formalism could be referring to anything physically real' (32) and ignores the subtleties of his views on complementarity — she takes more time with Heisenberg, presenting him as a visionary looking toward 'a new kind of metaphysical reality' (36). Indeed, Kastner characterizes her own goal as picking up where Heisenberg left off in his late-career philosophical writings.

Chapters 3, 4, 7 and 8 constitute the heart of the book; as noted earlier, Chapters 5 and 6 are technical digressions and can almost be treated as extended footnotes. Here Kastner introduces (Chapter 3), considerably extends (Chapter 4), and explores the ontological consequences of (Chapters 7 and 8) John Cramer's 'transactional' interpretation of quantum mechanics. Like Hugh Everett's 'relative state' interpretation — known following its reformulation by Bryce DeWitt as the 'many-worlds' interpretation — the transactional interpretation is based not on philosophical presuppositions but on a close and literal reading of the mathematical formalism. It begins by noting that quantum theory, like classical electrodynamics, is time-symmetric. Schrödinger's wave equation can, therefore, be written in two versions, one describing a wave — a quantum state — propagating forward in time, and the other describing a quantum state propagating backward in time. When forward- and backward-propagating quantum states overlap in phase — peak-to-peak and trough-to-trough — they reinforce each other in the same way that light, sound, or even ocean waves do. Cramer called such an overlap a 'transaction' between the forward- and backward-propagating waves, and proposed that such transactions, not quantum states themselves, are what can be measured as spots on a photographic film or clicks of a Geiger counter. In Cramer's picture, radioactive sources or hot filaments are 'emitters' of forward-propagating or 'offer' waves, while photographic films or Geiger counters are 'absorbers' that generate backward-propagating or 'confirmation' waves. A general-purpose absorber like a Geiger counter generates confirmation waves for all of the events it can detect; the offer waves that are actually encountered determine what events are actually detected.

As Kastner emphasizes, the transactional interpretation replaces the observer-induced 'collapse of the wavefunction' posited by the Copenhagen interpretation with a purely physical process. It therefore offers the philosophically-attractive possibility of a fully observer-independent and therefore objective quantum theory, in which determinate 'classical' outcomes are not the result of someone looking, but simply the result of Nature going about her business. It does so, moreover, in a way that explains the mathematical structure of the Born rule, the rule for calculating the probabilities of different outcomes for a defined measurement of a defined quantum state. As the Born rule must be assumed as an axiom in Copenhagen quantum theory, this is a substantial conceptual advance. Just as importantly for Kastner, treating absorption as a '*real physical process*' (55, emphasis in original) allows events — i.e. transactions — in which a quantum state is absorbed to be treated as unambiguously actual. It thus allows the transactional interpretation to be fully realist about the classical world of ordinary experience. Being fully realist about both quantum and classical worlds is the Holy Grail sought by Einstein, Bohm and Bell. Standard interpretations of quantum theory fall well short of this goal, and hence face — whether they admit it or not — a stark choice of viewing either the classical or the quantum world as essentially illusory.[1]

The obvious question raised by the transactional interpretation is whether it is too good to be true. Demonstrating that it is not requires explaining what 'emission' and 'absorption' of quantum states amount to, and hence what the transition from 'possibility' to 'actuality' really is. To do this, Kastner moves from the language of ordi-

[1] Landsman, N. P. 2007. Between classical and quantum. In *Handbook of the Philosophy of Science: Philosophy of Physics*, ed. by J. Butterfield and J. Earman, 417-553. Elsevier.

nary, nonrelativistic quantum theory to that of relativistic quantum field theory. Here ‘emission’ and ‘absorption’ become the actions of creation and destruction operators on a quantum field, the field corresponding to the quantum states of interest. This is a bold move: quantum field theory is generally thought of as a theory of elementary entities such as electrons or quarks, not as a general approach to quantum states. By glossing emission and absorption as field-theoretic creation and destruction, Kastner is implicitly proposing that every quantum state can be viewed as a field excitation, and hence that every collection of degrees of freedom amenable to quantum-theoretic description can be considered as a quantum field. Just how bold this identification is becomes clear in the discussion surrounding Figure 4.1, which shows a man observing a table. Kastner’s caption reads: ‘Macroscopic objects are perceived via transactions between offer waves emitted by components of the object and confirmations generated by absorbers in our sense organs’ (71). What ‘components’? The surrounding discussion suggests that the ‘components’ are atoms. Is Kastner referring to creation and destruction operators for generic quantum states of electrons, or is she suggesting that there are specific quantum fields for carbon, oxygen, iron and so forth? Or are the relevant ‘components’ themselves macroscopic? Is Kastner suggesting that there are quantum fields for tabletops?

The answer to these questions does not come until almost 100 pages later, and is developed in Kantian terms. The observer’s experience of the table as a macroscopic object, and hence the intersubjectively-confirmable existence of an ‘empirical world’ is taken for granted. As a realist, Kastner must postulate something external to the observer — an ‘object in itself’ — that generates the experience. This entity is defined by contrast to the ‘empirical’ object: ‘the ‘object in itself’ is precisely *that aspect of the real object which is not perceived*’ (162, emphasis in original). Kastner continues: ‘the ‘object in itself’ can be considered to be the offer wave(s) giving rise to possible transactions establishing the appearances of the object’; such objects ‘do not live in spacetime and can be considered a kind of abstract but physically potent information’ (162). The answer, then, is that the quantum field in question is a quantum field of information, information that observers experience as a table. This idea that quantum theory is at bottom a theory of information has become commonplace in the past decade, but is advocated primarily by instrumentalists, not realists. Kastner generally avoids the 'quantum information' vocabulary in *The Transactional Interpretation of Quantum Mechanics*, possibly to avoid its typically instrumentalist connotations. She therefore misses what would seem to be an obvious inference: if offer waves are a kind of physically potent information, confirmation waves must be a kind of physically potent information, too. Confirmation waves, however, originate in the observer; if they are physically potent information, they are physically potent information that the observer brings to the table. Kastner treats observers in the standard Einsteinian way, simply as points of view on a physical situation. The potential consequences of an observer contributing physically potent information to an observation are never explored.

What Kastner does explore, in Chapters 7 and 8, is the idea that offer waves — and hence confirmation waves also — do not 'live' in spacetime. It is this idea that principally distinguishes her 'possibilist' extension of the transactional interpretation from Cramer's original. From a purely formal perspective, this is obvious: quantum states have always 'lived' in Hilbert space, not in spacetime. From a physical perspective, however, explicitly banishing quantum states from spacetime is refreshing; the confusion of entanglement with faster-than-light communication, for example, is only a confusion if an entangled quantum state is imagined to be localized in spacetime. As Kastner points out, banishing quantum states — in the form of offer and confirmation waves — from spacetime immediately banishes 'particles' as well. If an electron, for example, is a 'quantum' particle, it cannot be localized; this is the fundamental lesson of quantum field theory. Banishing quantum states from spacetime also banishes causality; Kastner agrees with Hume and Russell that 'causality is *not* an ontological feature of the world' (166, emphasis in original). It is just an inference from the high probabilities of certain transactions.

Where then does spacetime come from? The standard position in quantum theory is a kind of embarrassed silence; position and hence spatial location is a quantum-theoretic observable, whereas time is not. Kastner's position on this question, outlined in Chapter 8, is both interesting and problematic. She starts with the idea of an observer — again, an Einsteinian point of view — for whom 'the past' is the backward-facing light cone. This past is populated by empirical observations: actualized transactions. The future, however, is not actualized; it is filled, to revert briefly to Cramer's original conception, with offer waves that have not yet arrived. Kastner calls this 'space' of unactualized possibilities 'prespacetime'; as a container for quantum states, it has the properties of Hilbert space. What connects prespacetime to spacetime is what happens at the instant that defines the present: absorption, and hence the actualization of some transaction. Kastner presents this idea with an analogy: the spacetime past is like a knitted fabric that '*continually falls away from us*' (176, emphasis in original), i.e. from our intersubjective and hence approximate present.

The idea that empirical, spacetime-bound reality is created out of a prespacetime of physically-real possibilities by a physical process of absorption allows Kastner to tackle philosophical riddles from the origin of time to the possibility of free will. In doing so, however, Kastner sidesteps the issue that has been lurking in the background since Figure 4.1: why a table? How does the physical process of absorption assign collective properties — or classical, collective degrees of freedom like center-of-mass position — to zillions of elementary emitters? Or are the zillions of elementary emitters somehow preorganized into macroscopic collectives like tables? Where, in other words, does classical information, the kind that can be written down in laboratory notebooks, come from? In particular, where does classical information about gross, macroscopic properties — the mass of the Earth, for example, or the 3-dimensional structure of your laptop computer — come from? Saying that this information is generated by the same process that generates space and time is helpful and possibly correct, but it is not enough. What is most perplexing about the story of Schrödinger's cat, after all, is not how the cat's quantum state collapsed, but how Schrödinger could ever have found his cat, let alone confined it to a box, before its state collapsed.

The only answer to this question seems to be spontaneous symmetry breaking, the inscrutable stochastic process that determines, in Kastner's often-repeated example, why the 24-pointed splash of a droplet of milk is oriented this way or that way with respect to a coordinate system in the plane of the splash. Spontaneous symmetry breaking is commonly evoked in association with the big bang — or these days, in association with inflation — but there is a long explanatory road to travel between the big bang and the existence of macroscopic entities like cats or tables. The usual story told along this road is evolutionary, but getting this story going requires classical information, in particular, classical information about entities that maintain their identities through time. Quantum theory does not provide us with this information, and the extension to relativistic quantum field theory does not help in this regard.

Kastner closes *The Transactional Interpretation of Quantum Mechanics* with a diagnosis: 'it is the omission of the back-reaction (i.e. absorption) which gives rise to the notorious intractability of the measurement problem' (204, parenthetical added). This is surely true: there are no detection events until whatever is being detected interacts with whatever is doing the detecting, and as Newton told us, interactions go both ways. Characterizing the back-reaction of the absorber, however, requires characterizing the absorber itself. It is this question — what is the absorber, or more poetically, what is the observer — that lies at the heart of the measurement problem. Despite over 80 years of effort, interpretations of quantum theory have yet to answer it.

The main virtue of a philosophical book, however, is to raise questions, not to answer them. *The Transactional Interpretation of Quantum Mechanics* raises many questions, and raises them forcefully and well. Anyone interested in the thorny questions of possibility and actuality will find it intriguing, and with some study, perhaps inspiring.

Chris Fields
Caunes Minervois, France
fieldsres@gmail.com

**Freedom of the Will: A Conditional Analysis**, by Ferenc Huoranszki. New York: Routledge, 2011, 208 pages.

BIBLID [0873-626X (2013) 37; pp. 368-374]

Huoranszki's *Freedom of the Will* is a book length defence of classical compatibilism, a position which affirms, as one condition on an agent's freedom, that the agent possess the ability to do otherwise. The book is a rewarding read and contains useful commentary on a number of long standing debates surrounding moral responsibility. In Part 1 of the book Huoranszki investigates how a number of foundational issues in action theory relate to the issue of responsibility. High points include an argument to the effect that the abilities relevant to free will must be extrinsic (37-41) and examples purporting to demonstrate that intentional control is neither necessary nor sufficient for responsibility (44-47). Part 2 applies the account developed in Part 1 to issues such as rationality, autonomy, reasons, and self-determination. Particularly interesting here is the claim that sensitivity to reasons is a condition, not on free will, but on autonomy (100-108).

Huoranszki's primary concern is to present a conditional analysis of free will which he does in chapter 4. The standard way of formulating such accounts is as follows: *S can φ iff S would φ if S chose to φ*. This kind of conditional analysis renders free will compatible with determinism, but, as van Inwagen has taught us (Peter van Inwagen, *An essay on free will*, Oxford, 1983, 121), the conditional analysis of 'can' is really nothing other than the compatibilist's central premiss. To establish compatibilism over and against incompatibilism, therefore, it is not enough to present a conditional analysis of 'can' if at the same time there are compelling arguments for incompatibilism. Huoranszki agrees with this point (57), which is why, after presenting his broad framework in chapter 1, chapter 2 kicks off with an analysis and attempted refutation of the consequence argument. Huoranszki's discussion of the consequence argument is interesting because he does not spend much time on the usual intricacies concerning the various transfer principles, but instead aims to undermine the intuitive force of the argument.

Huoranszki draws a distinction between the concept of *determinism* as used in the argument and a concept of *determination* which uncontroversially threatens freedom (29). The former is *global*, in that it refers to states of the whole universe, and *abstract*, in that it does not refer to any particular laws. In addition, the consequence argument's assertion that the 'propositions expressing any physical state of the universe at one instant and propositions expressing the totality of laws of nature imply propositions about the physical states of the universe at all other instances' is itself a *consequence* of determinism, rather than a part of the thesis (29). Huoranszki's point seems to be that the determining going on in the consequence argument is all very theoretical and far removed from our ordinary lives. To further trivialise this determining he draws a parallel between the implications in the consequence argument and the obviously banal implications that hold between other kinds of propositions. A proposition about someone's being a bachelor *implies* a proposition about someone's being unmarried, and — so the thought seems to go — *in a similarly trivial manner* a proposition about the past *implies* a proposition about the future, if determinism is true.

What the incompatibilist gets right, Huoranszki says, is a hostility towards *determination*: the idea that something *local* has caused someone to do something (28). This is what 'determinism' means in the context of ordinary language, and such local determination — psychological, social, or genetic — would indeed be incompatible with freedom. Such notions have been shown false, Huoranszki thinks. The incompatibilist argument gains traction only by trading on the ambiguities in word determinism, and when 'we realise how abstract and global the notion of determinism involved in the consequence argument is, it is already less clear how that sort of determinism can deprive us of our free will' (29).

It is true, of course, that the consequence argument employs entailments between propositions, and also true that the content of those propositions is general and abstract in the way Huoranszki describes. But it is hard to see how that is anything but a virtue: the argument applies to every agent at every time, and it is not held hostage to any empirical discoveries. It is also hard to fathom why the implications highlighted by the consequence argument would be rendered irrelevant to issues of freedom because of the 'similarity' they bear to the implications that hold between propositions about *bachelors* and *being married*. What underlies these implications is entirely different: in one it is facts about causation and natural laws, in another it is conventions of meaning. If the implications highlighted by the consequence argument were rendered trivial by this similarity, no argument in philosophy would be safe.

One of the most thought provoking portions of the book is Huoranszki's discussion of the nature of abilities and their connection to responsibility. It is common in the philosophy of action to think of responsibility deriving from some set of basic actions which are often thought to be one's physical bodily movements. On this view, responsibility 'flows' from an agent's bodily movements to their complex actions. We are directly responsible for basic bodily actions (moving my finger) and we are derivatively responsible for complex actions and their consequences (turning on the light). Huoranszki thinks this view is mistaken. Although it might be correct to say of any complex action that we do it *by* doing some basic physical action, this by-locution is not the by-locution which connects cases of direct and derivative responsibility. Responsibility, in other words, does not originate in basic actions.

Consider the following example, which Huoranszki uses to argue for this view: suppose, intending to insult someone, I say something rude to them. My bodily movement here is a set of tongue and mouth movements, but it is highly plausible that what I'm directly responsible for is saying something rude. This is because I have no conceptual representation of the mouth movements qua mouth movements and so I *could not* make the mouth movements directly. The only way for me to make those mouth movements is by *saying those words*. What this shows is that responsibility for complex actions such as insults cannot be derivative, being built up from the responsibility from basic actions which compose them, because the basic actions which compose them cannot be performed independently of the complex action (39).

Huoranszki takes this example, and a number like it, to show that the actions we hold each other directly responsible for are specified with reference to extrinsic results, and from this he concludes that the abilities relevant to free will must also be specified extrinsically. This idea forms a theme which runs through the whole book

(32, 36-44, 62, 84-89). Despite being clear about the broad outline, however, Huoranszki's account does not fill in as many of the details as one might like. Here are some of the questions that need to be asked of the thesis:

(1) What is the scope of the thesis?
(2) What is the sense of extrinsic in play?
(3) Do the examples provided support it?

Consider question (1). At the close of chapter 2 Huoranszki says that 'those actions for which we're responsible … are *almost never* intrinsically identified' (32), but a few pages later we are told that 'the types of action for which agents are responsible *must* be extrinsically identified' (36). This might be a minor issue, but it is not entirely inconsequential: allowing exceptions would preclude the set of abilities relevant to free will being of a unified metaphysical kind, and this might preclude certain kinds of explanation for that class of ability. (A simple example suggests that we do need to allow exceptions: e.g., one can be directly responsible for the basic bodily movement of *stepping* onto a patch of grass in the vicinity of a 'Do not step on the grass' sign.)

Question (2) is more important. What is it to be extrinsically specified? Huoranszki says that abilities are 'extrinsic in the sense that the ascription of such abilities is sensitive to conditions that lie beyond the agent's body' (62), and elsewhere he is clear that abilities can be lost even when the agent undergoes 'no internal change' (85). This suggests that 'extrinsic' is to be understood as 'external' as opposed to 'relational.' And external circumstances feature centrally in Huoranszki's account of abilities: he eschews any ability/opportunity distinction as a useful way of thinking about the free will problem (31-2), saying instead that we must pay attention to when opportunities affect the possession of abilities.

But *which* extrinsic conditions affect whether an agent possesses an ability? Huoranszki says that abilities need to be *maximally specific* (24-5, 84). We are not told what this means. The implication is that a specific ability, in contrast to a general ability, will be sensitive to (more?) features of the environment. If we take *maximal* at face value we might think such a specific ability will be sensitive to *all* the extrinsic conditions. But his discussion of Frankfurt style-cases shows that this is not what he means: the presence of the intervener *does not* remove the agent's ability to perform the action in question (82-9). The intervener — an actual, extrinsic factor — is not to be taken into account when assessing the agent's ability. (Interveners are sometimes described as *counterfactual*. This is unfortunate and leads to many misunderstandings. It is only the intervener's intervention which might properly be described as counterfactual). Compare the above with Huoranszki's judgement of Locke's man unknowingly put in a locked room case. Here Huoranszki says that the man *does not* stay in the room of his own free will (32); the locked door removes the man's ability to leave. This is problematic for Huoranszki for these cases appear analogous. In both we have an extrinsic factor which blocks the agent doing something. Locke's man *will* remain in the room and in the Frankfurt-style case the result the intervener wants *will* occur. In neither case is there anything the agent can do to avoid the end result. Not only, therefore, do we lack a principled way to judge which extrinsic circumstances to use when assessing an agent's ability, we seem to have a set of analogous cases for which Huoranszki has given differing judgements.

Consider now question (3): does the example support the thesis? What seems immediately clear is that the speech example is a strong counter-example against the basic/non-basic action distinction as drawn by Danto. This is because the agent cannot conceptualise the mouth movements but they can conceptualise insult. But this does not show that the agent's *possession of the ability* depends on extrinsic facts about the agent's environment, it shows only that the agent's representation needs to be about some external state of affairs.

Let us move now to core of the account, the conditional analysis of free will, Huoranszki's version of which runs as follows (66):

> S's will is free in the sense of having the ability to perform an actually unperformed action A at t iff S would have done A, if
>
> (i) S had chosen so, and
>
> (ii) had not changed with respect to her ability to perform A at t, and

(iii) had not changed with respect to her ability to make a choice about whether or not to perform A at t.

The antecedent of the conditional contains three conditions, in contrast to the usual one. As Huoranszki notes (55), however, Moore's own account of free will included two conditions over and above the simple conditional that it is sometimes thought to consist in. Huoranszki doesn't think Moore's further conditions were adequate, but takes himself to be improving on the model supplied by Moore. How does Huoranszki's account fare? The objection deemed decisive in burying the simple conditional analysis was articulated with great clarity by Lehrer (Keith Lehrer, 'Cans without ifs', *Analysis* 29 (1), 1968: 32), and the key thought is this: the truth of the conditional is not sufficient for the truth of the ability ascription. Lehrer's example has become prodigious in the literature: suppose I am presented with a bowl of red candy, and while I might like candy in general, and am not paralysed, I have a pathological aversion to taking one of these candy because they remind me of drops of blood. The following two things are true:

I cannot take a red candy.

I would take one, if I choose to.

This suffices to show that 'I can' cannot mean 'I would …, if …'. Huoranszki's clause (iii), which affirms not just that the agent is unchanged in their ability state but that they do indeed possess the ability, is introduced explicitly to address this problem. The clause works by conceding Lehrer's point that 'would if I choose' only stands a chance of being part of the correct analysis of 'can' if 'I can choose' is also included.

Crucially, clause (iii) includes reference to the agent's *ability to choose*. Another ability. And one which cannot be analysed using the above account on pain of circularity. Not that Huoranszki advocates such a thing. He maintains that the ability to choose is an entirely different kind of ability because *making a choice is not an action* and the control we have over our choices is different to the control we have over our actions (47, 51).

This is what allows Huoranszki to resist any charge of circularity. In bottoming out in a (putatively) non-actional component, his account is akin to that of Davidson's, who cited a belief/desire pair as the antecedent of the conditional. But whereas beliefs and desires are uncontroversially non-actional, choices are far from being so. The key, Huoranszki thinks, is to understand *choice* as referring not to a stretch of deliberation but to the end result, the 'coming to a practical conclusion' (51). This is non-actional because agents do not control the results of choices in the same way they control the actions they perform. In support of this Huoranszki invokes Locke's famous point, namely, that once the possibility of an action has occurred to us, our choice concerning it cannot be free: we can choose whether or not to do it, but we cannot choose whether or not to choose about it. These claims are contentious, but concede them for argument's sake. Huoranszki's account still has a major problem because alongside the non-actional aspect of choice he is keen to countenance the actional aspect: he speaks of deliberation as an *act* (51), and he refers to the *activity of choice making* as both intentional and voluntary (52-3). Whatever we call this actional component of choice, Huoranszki's account needs to apply to it. This will either make the account circular or, if it is argued that the conditions of responsibility for this kind of action are for some reason different (which itself risks being ad hoc), the account will be incomplete.

Despite the above problems, Huoranszki's account is a strong defence of compatibilism. The high level of detail in many sections of the book will repay careful study.

Simon Kittle
Department of Philosophy
University of Sheffield
45 Victoria Street
Sheffield
S3 7QB
United Kingdom
simon@kittle.co.uk

**The Origins of Grammar**, by James R. Hurford. Oxford: Oxford University Press, 2012, 808 pages.

BIBLID [0873-626X (2013) 37; pp. 375-381]

James R. Hurford's *The Origins of Grammar*, another title adding to the ever-increasing literature on the evolution of language, happens to be the second of a two part collection touching on many other issues. The first, titled *The Origins of Meaning* and published in 2007, focused on the evolution of conceptual thought and communication from the perspective of animal cognition, setting the stage for the evolution of language that is tackled in the second volume. It will do us well, then, to summarise the prior volume before getting to grips with the second, much longer, one.

Evenly divided in two parts, the 2007 book starts by analysing the nature of animals' conceptual representation systems, arguing that differences with human conceptual systems are in most cases a matter of degree rather than kind. This is of course a rather controversial claim, but Hurford does offer evidence for the proposition that animal cognition is underlain by a rather rich conceptual structure; in particular, he argues that one can find examples of predication and propositional structure, reference and deixis, and some sort of episodic memory in animals' cognition. The second part of the book focuses on animal communication, a phenomenon Hurford characterises as doing things to each other dyadically, in contrast to the triadic relationships that typically arise in human communication between speakers, hearers and whatever is being discussed. Despite the reputed differences, these phenomena, Hurford tells us, constitute the 'seeds' of his evolutionary story: these animal abilities constitute the precursors to human cognition and language. How one goes from that to the full glory of human grammar is of course what needs to be explained; hence, volume two.

Alas, that is a long way out in *The Origins of Grammar*, as the relevant material, the 'what happened' of part 3, only comes in on page 481. Before that, part 1 discusses the 'twin evolutionary platforms' of language — animal syntax and lexicon —, whilst part 2 offers a crash-course in linguistic theory (some 300 hundred pages of it, though). Less roughly, part 1 is divided into two chapters, the first of which has a long look at the structural features of various animal communication systems, showing that none of these systems approximate human language because, first, they do not exhibit the right expressive power (at least in terms of the formal languages and grammars of the Chomsky Hierarchy) and, more importantly, they all lack a 'complex semantically compositional syntax' (21), perhaps the key feature of human language. Chapter 2, in turn, focuses on the 'development of a shared system of conventionalized symbols' (153) — a lexicon — the origin of which is to be found, it is suggested, in animals' gestures and some sort of process employing sound symbolism and/ or synaesthesia (127). Moving on, the tutorial starts by nominating Construction Grammar (CG) as the linguistic framework of choice, supposedly for being more compatible with the gradualist account of evolution Hurford adopts throughout (Chapter 3, 177-80). It is finally at this point that we are told that what is at stake here is an account of the evolution of linguistic knowledge, that constitutive part of a speaker's linguistic capacity which explains overt behaviour (207-8), a way of putting things that respects the well-known competence-performance distinction (with some modifications). In turn, Chapter 4 lists, and describes to some length, those features of language Hurford claims to be universal (among others, a massive store of symbols and the 'constructions' CG posits — form and function pairs that speakers store, supposedly resulting in a syntax-lexicon continuum). Finally, Chapter 5 argues that languages vary in complexity at different levels, the most important of which involves whether a language uses inflectional morphology, function words, single-valence verbs, and serial verb constructions (459). Crucially, these features relate to some of the pre-syntactic properties that Hurford is to focus on in part 3: topic-focus word order and concatenation (a case of non-hierarchy). The latter are present, we are told, in pidgins, new sign languages, and early child language, the linguistic systems Hurford feels to be most informative of how language actually evolved. Finally, Chapter 6 opens part 3 by summarising and centring the 'pre-existing platform' discussed both in the 2007 book and in the first two-thirds of this book (basically: rich conceptual representations, massive storage, and some sort of syntax). These features of animal cognition, it will be recalled, constitute the 'minimal seeds' of the language faculty, properties that differ from those of modern humans' only in degree, discounting, Hurford thinks, the need for 'some incomprehensible process [to] bridge the gap' between animal and human cognition (537). Viewed this way, and with a little help from CG, the combinatoriality so often stressed of human language is 'itself no big deal' (ibid.), given that all is needed to account for it is a massive store of constructions that combine with each other, template-like. Of course, we are still owed an account of what happened in the 4 million years that separate our species from the pre-human ancestors, and Chapter 7 breaks the ice by offering some speculative ideas regarding how the first arbitrary symbols came to be learned (a different question to the origin of symbols more generally, which Hurford connected, it will be recalled, to animals' gestures and sound symbolism). Such a phenomenon could have come about, we are told, by 'the combined effects of increased group size, increased cooperation within groups, increased trust, and shared intentionality' (563), making our species enter a 'symbolic niche' in which exchanging messages for cooperation would have improved the (evolutionary) interests of both the individual and the group (564-5). From here, an ever more complex system was inevitable, it seems, as more complex messages would improve the chances of survival. There must have been, then, a transition from proposition-expressing one-word utterances (596 ff.) to two-word concatenations (606), the latter composed of a symbol expressing 'what is most urgent to convey' followed by a symbol that states 'what is next uppermost in [the] mind' (607) — think of constructions such as *Mommy sock*, typical of toddlers. Apparently, such deictic + predicating word constructions are typical of trained apes, pidgins, infants learning the ambient language, child deaf home-signers and creators of sign languages (620). How do children, in particular, go from this to multi-word combinations, though? By employing a synthetic, putting-things-together type of process, one typical of the theory of language acquisition CG favours — that is, a case of acquiring constructions gradually, (638-9). And since language acquisition is 'the most promising guide' to understanding language evolution (590), this is the most plausible scenario for the origins of grammar. The book ends with a chapter on grammaticalisation, a robust linguistic phenomenon according to which the 'effects of frequent use ... become entrenched as part of the learned structure of a language' (646). It is here put to use to claim that topic-focus pairs — the central two-word combinations said to be present in all languages and of an early appearance in language acquisition and therefore in language evolution — derived the difference between nouns and verbs, and eventually that of subjects and predicates, by way of a grammaticalisation process (648 ff).

So much for a lengthy description of the book; in what follows, I aim to evaluate it along two dimensions: conceptually and empirically. That is to say, I first intend to discuss the underlying assumptions and then proceed to assess the empirical case for the scenario Hurford proposes.

Regarding the first set of issues, it is noteworthy that for an evolutionary book there is actually very little about the theory of evolution itself. As mentioned, Hurford adopts a gradualist account of the evolution of grammar, but we are hardly offered any details regarding what considerations precisely will be pertinent. All we are told is that such a stand is more plausible than 'saltations to syntax' (180), which may well be true, but given that it is not explained exactly why, the election is somewhat unprincipled. Further, it will be recalled that CG was adopted on account of being more compatible with a gradualist approach, but no more than that was offered as a way of justification, which is unfortunate for a number of reasons.

In recent years, there has been a convergence of various grammatical formalisms (minimalist grammar, tree-adjoining grammar, combinatory categorial grammar, etc.) regarding both the set of primitive computational operations they postulate (merge, adjunction/substitution, and composition — the basic rules of the aforementioned formalisms — have been shown to be roughly equivalent) and the expressive power they must model (natural language is said to be mildly context-sensitive; see Edward Stabler 'Computational perspectives on minimalism', in The Oxford Handbook of Linguistic Minimalism, ed. by Cedric Boeckx , Oxford, 2011, for details). Noticeably, CG is not part of this convergence in either respect, which is not an insignificant shortcoming, given that capturing the right expressive power of language ought to be regarded as a lower bound that any grammar must meet if it is to be considered adequate at all. This point goes unmentioned in Hurford's book, despite devoting ample space to the issue of expressive power in Chapter 1, even discussing some of the grammatical formalisms I have mentioned.

Further, there is a wide range of syntactic data that CG quite simply fails to account for (see Jeffrey Lidz & Alexander Williams, 'Constructions on holidays', in Cognitive Linguistics, 20, 2009), casting doubt on the proposal to reduce all linguistic knowledge to form-function pairs (constructions). As Lidz & Williams point out, all sort of theories accommodate the notion of constructions, especially for the data most easily accounted for in these terms (such as idioms and argument structure), but there is a long way from that to the conclusion that linguistic knowledge in toto should be so described — and even more far-fetched is the resulting lexicon-syntax continuum that surely makes Hurford underestimate the importance of the syntactic engine.

Further still, Hurford's account is hostage to constructionist theories of language acquisition, particularly that of Michael Tomasello that he so approvingly cites (589 ff), and this is a crucial failing. Indeed, Hurford sees Tomasello's (2006, cited therein) 'overview of the course of child's grammar acquisition, couched in terms of acquiring constructions' as a theory that 'works' (589), and thereby goes his theory of language evolution. That Tomasello's theory of language acquisition 'works' is bound to surprise many a linguist; alas, it does not.

According to Tomasello's usage-based account (2006, loc. cit.), language acquisition proper starts at age 9-12 months, once the intention-reading skills of children develop, allowing them to recognise the communicative intentions behind the noise parents are directing at them. At this stage, children start producing holophrases — one-word utterances that describe an entire experiential scene —, the product of imitation, intention-reading, and cultural learning processes (ibid., 268). Building upon that, and by employing these processes further, 18-month-olds start combining words in pairs, thereby partitioning a scene into units (ibid., 269). These pairs gradually develop into pivot schemas, structures in which an element organises a whole utterance (such as *More X*; ibid., 270). A product of some sort of schematisation process, these schemas eventually become item-based constructions (by age 24 months and onwards), the first manifestation of word order and participant roles; that is, of syntax (as in 'pushee-pusher' pairs, which correspond to Hurford's two-word constructions; ibid., 270-2). By age 36 months, children start employing pattern-finding abilities, most notably that of analogy, to form the first abstract constructions, the result of generalising across many dozen item-based constructions (ibid., 279 ff). As a child grows older, it constructs ever more abstract linguistic knowledge thanks to various processes, such as entrenchment and pre-emption (the former probably operative before the age of 3), and functional based distributional analyses (ibid., 287 ff).That is not all; all these processes are supposed to work in conjunction with syntactic operations such as 'adding on', 'filling in', and 'cut and paste' (ibid., 291).

Unfortunately, none of these constructs, be they syntactic operations or the learning processes that Tomasello proposes (intention-reading, imitative learning, schematisation, etc.) are actually properly described, or indeed explained. In fact, nowhere is it shown, and this is especially true in the case of the very important process of analogy, how children actually do such things — if indeed they do (this much is admitted by Tomasello himself when he recognises that there is no 'systematic research' into how children align verb meanings 'in making linguistic analogies across constructions'; ibid., 285). Put simply, there is literally no demonstration of how children's linguistic knowledge moves from one stage to the other by employing the operations and processes postulated.

In any case, the main problem with Tomasello's take on things is that the expressions that children produce are taken to be a faithful representation of the structures that children mentally represent — i.e., of their grammar — but that is entirely unwarranted. Indeed, it is widely accepted within modern linguistic theory that language is mainly an internal phenomenon not always transparently manifested in overt behaviour. Relevant to our purposes here, two-year-olds are said to produce 'telegraphic speech', but this cannot be what they represent in their minds, for they are perfectly capable of understanding what is said to them in normal speech — their grammar allows it, suggesting a much richer structure than what is being produced (See Letitia Naigles, 'Form is easy, meaning is hard', Cognition, 86, 2002, 175-6, for references). Tomasello would have you believe that these very two-year-olds do not possess much abstract grammatical knowledge, given that this is not evidenced in the experimental tasks he uses (all geared towards eliciting verbal responses), but comprehension tasks clearly show that these children understand many complex linguistic structures way before they can actually produce them (Yael Gertner, Cynthia Fisher & Julie Eisengart, 'Learning words and rules', Psychological Science, 17, 2006).

In sum, the linguistic behaviour of children does not correspond to their actual, internalised knowledge; nor does this knowledge go through the stages Tomasello has outlined, only their productions do — and mutatis mutandis for child deaf homesigners and creators of new languages, signed or otherwise. Thus, there does not appear to be a synthetic gradual process in which children go from two-word constructions to full syntax. If Tomasello's theory of language acquisition is the main motivation for Hurford's account of the evolution of language, then his own story can itself be neither coherent nor true.

So much for Hurford's conceptual assumptions and empirical details, then. I shall put an end to this review by making one final point. For a story of how language evolved, there is actually very little about the mental organisation underlying the language faculty. It is widely-accepted that language is that system of the mind that generates sound-meaning pairs, and this implicates a number of different components; at the very least, the sensori-motor systems, a phonology component, whatever semantic/conceptual structure participates, a syntactic engine and lexical items (roughly, bundles of phonological, syntactic and maybe semantic features that come together to form what we call words). Note that it is the coming together of these probably independent systems that gives you language. Thus, it is the emergence of such mental organisation, I should think, that an evolutionary account of language ought to elucidate.

David J. Lobina
University of Oxford
Faculty of Philosophy
Radcliffe Humanities
Radcliffe Observatory Quarter
Woodstock Road
Oxford, OX2 6GG, England
david.lobina@philosophy.ox.ac.uk