Received: 14 September 2018 | Revised: 1 February 2019 | Accepted: 27 April 2019

DOI: 10.1111/rati.12241

ORIGINAL ARTICLE

WILEY

Dangerous Liaisons

Roberto Horácio de Sá Pereira

Department of Philosophy, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Correspondence
Roberto Horácio de Sá Pereira, Department of Philosophy, Universidade Federal do Rio de Janeiro, Largo de São Francisco número 1, Rio de Janeiro 22610-180, Brazil.
Email: robertohsp@gmail.com

Funding information
Research for this article was funded by CNPQ, Conselho Nacional de Desenvolvimento Científico e Tecnológico

Abstract

In this paper I take side on externalist incompatibilism. However, I intend to radicalize the position. First, based on my criticism of Burge's anaphoric proposal, I argue that there is no reasoning-transparency: the reasoner is blind to the reasoning he is performing. Second, assuming a global form of content-externalism, I argue that "in the head" are only logical and formal abilities. That is what I call "bite the bullet and swallow it too."

KEYWORDS
content-externalism, content-transparency, dangerous reasonings, rationality failure, safe reasoning

1 | INTRODUCTION

The attack on content-externalism on the grounds that it is incompatible with content-transparency has been much debated over the decades. The first line of argumentation claims that content-externalism undermines the a priori authoritative self-knowledge or the authority of the first-person. According to content-externalism, the content of mental states is determined, in part, by the non-intentional relations that the subject has with her/his environment. Given this, the content of the subject's mental state could change without her/his knowledge under the assumption that the subject undergoes undetectable environmental changes, say, from Earth to Twin-Earth and vice-versa. Thus, if content-externalism is right, one can never authoritatively know a priori whether one is now thinking about *water* or *twater*.

According to Burge, however, content-externalism is not a threat to cogito-like thoughts, that is, second-order contemporaneous thoughts about one's first-order thoughts. Without the need for any empirical enquiry or any epistemic self-justification, the introspective second-order thought "Inherits," so to speak, the content of the first-order thought by embedding it. If some Earthian's first-order thought "water quenches thirst" refers to H_2O rather than to XYZ, his introspective second-order thought "I know that I am thinking that water quenches thirst" also refers to H_2O (instead of XYZ) because the first-order thought is embedded in the introspective second-order thought.

Yet, even if Burge's self-verification account of cogito-like thoughts is convincing, two additional key questions still remain open. The first question is whether under content-externalism the content can be preserved in slow-switching cases (introduced by Burge in the 1988 literature). Suppose that content-transparency fails and I am an accidental tourist that travels from Earth to Twin-Earth when I remember that Pavarotti once floated on Lake Taupo. What is the content of my memory? The second is a direct development of the first. Suppose now that the original content is preserved and I remember that the Earthian Pavarotti floated on the Earthian Lake Taupo. Yet, assuming assimilation to the dialect of twin Earthians, whenever I think from now on of Pavarotti I refer to Twin-Pavarotti rather than to the Earthian Pavarotti. The problem is that if the content is preserved I might from this infer that Twin-Pavarotti got wet, thereby making my inference infelicitous. Under content-externalism, reasoning is unsafe. This second question is what concerns me in this paper.

Let me provide you with big picture of the battlefield. We should discern at least three fundamental positions: (i) internist incompatibilism (Boghossian, 1992, 1994, 2011, 2012, 2015; McKinsey, 1991, 2002, 2018, et alia), (ii) externalist compatibilism (Burge, 1998; Garmendia, 2014; Ludlow, 1995, 1999; Recanati, 2012; Tye, 1998; Sainsbury and Tye, 2012, 2015), and (iii) externalist incompatibilism (Faria, 2009, 2010, 2015; Sorensen, 1998, et alia.). (i) Internalist incompatibilists argue that because of the lack of content-transparency reasoning becomes irrational in trivial cases. Let me call this the "fire the bullet" strategy. (ii) In contrast, externalist compatibilists maintain that the incompatibility is only apparent because they believe that there is always a way of showing that the reasoning is not irrational, as it prima facie appears to be. Let me call this the "exculpation" strategy: no one is to blame because there is no rationality-failure in the first place. (iii) Externalist incompatibilists reject content-transparency and assume what in the eyes of others seems absurd: our reasoning is really unsafe. Let me call this the "bite the bullet" strategy.

In this paper, I take side on externalist incompatibilism. However, I intend to radicalize the position. First, based on my criticism of Burge's anaphoric proposal, I argue that there is no reasoning-transparency: the reasoner is blind to the reasoning he is performing. Second, assuming a global form of content-externalism, I argue that "in the head" are only the logical and formal abilities. Thus, using Faria's words (2009), reasoning is *always* unsafe, and in Sorensen's words, *always* depends on (rational) luck (1998): the reasoner has to be in the right place and time.

How shall I proceed? First, I argue that the exculpation strategy is deemed to fail. I aim to show that its proponents fail to circumvent the "univocation"[1] and the equivocation fallacies by adding new implicit premises to the inferences (Garmendia, Recanati's case). Second, I argue that even when succeeding in circumventing the "univocation" and the equivocation fallacies, they fail to show that the reasoner knows which reasoning he is performing. Yet, the key question is: how can I attack internalist incompatibilism? What is modus pones for the "fire the bullet" strategy is modus tollens for the "bite the bullet" strategy, and I should provide an independent argument against content-transparency, the bone of contention between externalist incompatibilism and internalism incompatibilism. Given this, Sorensen (1998) appeals to vague ambiguities to refute transparency. Yet, I believe that such a strategy is far from convincing. For one thing, internalist incompatibilists may have no problem in conceding in the case of vague ambiguities and genuine indexicals that there is no transparency. Since their target is content-externalism, what they care about is transparency concerning the Twin-Earthable contents making the simplest inferences unsafe. Regarding this, I avoid any independent refutation of the transparency thesis in the case of the Twin-Earthable contents.

Finally, I argue where my "bite the bullet" strategy differs from Sorensen's in crucial points: I reject Sorensen's distinction between extrinsic and intrinsic fallacies and assume a global content-externalism. As I reject local content-externalism, I conclude that if reasoning is "formally" individuated by the logical form that it exemplifies, it is also "materially" individuated, by intentional and non-intentional relations that the reasoner bears to his/her environment and to his/her speech community.

This paper is conceived in seven sections. The next section is devoted to the exposition of Boghossian's transparency thesis and a critical analysis of Kripke's Paderewski case. Boghossian's arguments against the compatibility between content transparency and content-externalism are based on two well-known thought experiments:

[1]Sorensen's word. See 1998, p. 326.

Paderewski-Peter, stemming from Kripke,[2] and Switched-Peter, proposed by Burge in the so-called slow-switching cases.[3] In a nutshell, given the lack of content-transparency, Boghossian claims that Kripke's Paderewski-Peter falls prey to a blatant contradiction. Likewise, given the absence of content-transparency, Boghossian claims that Switched-Peter falls prey to a fallacy of equivocation in simple inferences.

The following section is devoted to presenting Burge's cohabitations thesis and his anaphoric solution to Boghossian's challenge. My claim here is not to refute Burge's claims that content can be preserved, but rather to show that his account reinforces my view that reasoning changes, regardless of his anaphoric account.

The short sections after that are devoted to a critical examination of the various exculpation strategies in slow-switching cases: the attempt to show that Switched-Peter does not fall prey to a fallacy of equivocation, Tye's replacement solution, and what I call the fusion solution. The penultimate section is devoted to a criticism of Sorensen's ingenious paper. In the conclusion section, I present my own view.

2 | THE VIOLATION OF THE TRANSPARENCY OF SAMENESS

Boghossian has mounted a sustained attack on content-externalism, on the grounds that it is inconsistent with content-transparency: either content-externalism is right and transparency must go, or transparency is right and content-externalism must go. Boghossian's standard formulation of content-transparency is this:

Introspective knowledge of comparative content. When our faculty of introspection is working normally, we can know a priori via introspection whether any two present, occurrent thoughts exercise the same or different concepts.

Boghossian breaks the Introspective knowledge of comparative content thesis up into two parts, namely the transparency of sameness and the transparency of difference. The first part is:

Transparency of sameness: If S's faculty of introspection is working normally, then if S occurrently thinks two thoughts that exercise the *same* concepts, he will be able to know that fact introspectively.

If S's faculty of introspection is working normally, then if S occurrently thinks two thoughts that exercise *the one same* concept, then S will be able to know that fact introspectively.[4] Now, with the exception of what Burge has called cogito-like thoughts, there is no question that content-externalism is incompatible with what Boghossian is calling the transparency of sameness here. Yet, according to Boghossian, the problem is that when we give up transparency, we open the door to failures in the simplest inference, where these failures are not supposed to happen. As Sorensen puts it, the converse of equivocation is "univocation" (1998, p. 326). So, the fallacy in question takes place because the reasoner treats a univocal expression as if it were ambiguous. In this section, I am concerned with Kripke's Paderewski-Peter fallacy of *univocation*. Following Kripke, let's assume that Peter hears Paderewski at a concert, and forms the belief:

[2]See Kripke, 1979.

[3]See Burge, 1998.

[4]Both claims echo Boghossian's formulation of 2011:

> (a) If two of a thinker's token thoughts possess the same content, then the thinker must be able to know a priori that they do; and (b) If two of a thinker's token thoughts possess distinct contents, then the thinker must be able to know a priori that they do. (2011, p. 1)

And his original formulation of 1994:

> The thesis of the epistemic transparency of content may be usefully broken up into two parts: (a) If two of a thinker's token thoughts possess the same content, then the thinker must be able to know a priori that they do; and (b) If two of a thinker's token thoughts possess distinct contents, then the thinker must be able to know a priori that they do. Call the first the thesis of the *transparency of sameness* and the second the thesis of the *transparency of difference*. (1994, p. 36)

1-Paderewski is a talented musician.

Later, Peter encounters Paderewski making a political speech at a rally. Now, having been told on good authority that no politicians have musical talent, Peter comes to believe that:

2-Paderewski is not a talented musician.[5]

The classical Fregean solution here is to assume that Peter associates two different modes of presentation (two concepts) with the same name "Paderewski" referring to the same person. Roughly, in 1 "Paderewski" presents PADEREWSKI as a pianist in a concert hall, while in 2 "Paderewski" presents PADEREWSKI as a politician. However, that solution is not available for Kripke's referentialism and for content-externalism here in general. First, according to Kripke's Millianism, names do not connote, but only denote; therefore, the meaning of a name is nothing but its reference. Thus, there is no way to excuse Paderewski-Peter for his contradiction by appealing to a Fregean Sense. Second, according to content-externalism, "meaning ain't in the head"[6]; that is, what determine the reference of "Paderewski" are the non-intentional relations of Paderewski-Peter with his environment. As Paderewski-Peter's concept PADEREWSKI is determined by means of his causal relations with *Paderewski*, regardless of whether he knows it a priori or not, he possesses just one concept PADEREWSKI.

Boghossian is right when he claims that Paderewski-Peter violates Transparency of sameness. However, the price to pay is very high: again, Paderewski-Peter commits a simple fallacy of univocation, where failures are not supposed to happen by holding that:

4-Paderewski is a talented musician & Paderewski is not a talented musician.

The putative blatant contradiction takes this simple form:

1-Ta.

2-~Ta.

4-Ta & ~Ta.

Peter could easily avoid the blatant contradiction if the transparency of sameness was met. If Paderewski-Peter knows that he has just one concept PADEREWSKI, he will never hold 4. Therefore, he comes to hold 4 precisely because he assumes that he possesses two homophonic concepts, PADERWSKI_1 and PADERWSKI_2.

The first exculpation strategy to address the fallacy of univocation is merely to assume that 3 below is an implicit premise in the reasoning from 1 to 4. Paderewski-Peter's contradiction expressed by 4 is excusable because he holds that:

3-PADEREWSKI held in 1 is not PADEREWSKI held in 2.

What we have in Paderewski-Peter's mind is the following:

1-Ta.

2-~Ta.

3-a in 1 (a_1) ≠ a in 2 (a_2).

5-Ta_1 & ~Ta_2.

However, there is little to recommend this exculpation. For if Paderewski-Peter does not know that he has just one concept PADEREWSKI, he does not know that he could possess two or more concepts, roughly PADERWSKI_1 and PADERWSKI_2, either. It seems much more plausible to assume that Paderewski-Peter's mental state is instead a non-propositional attitude.[7] The assumption is that Paderewski-Peter behaves "as if" he has two concepts, namely, PADERWSKI_1 and PADERWSKI_2, when in fact he has just one. Or, to put it in Boghossian's own words,

[5]Boghossian, 2015, p. 98.

[6]Putnam, 1975.

[7]See Sainsbury & Tye, 2015. To create the puzzle, it is not even necessary to assume that Paderewski-Peter believes that he has two PADEREWSKI concepts when in fact he has just one. All one needs is to assume that he behaves "as if he" has two concepts. See Sainsbury & Tye, 2015, p. 3. For one thing, we must be careful not to ascribe to Paderewski-Peter any meta-conceptual beliefs, for he might be in the relevant states even if they lack the concept CONCEPT, and so could not even formulate any meta-conceptual beliefs.

Paderewski-Peter non-propositionally "treats" the one same concept PADEREWSKI as if there were two.[8] Given this, the exculpation has to explain how Paderewski-Peter non-propositionally "treats" the one same concept PADEREWSKI as if there were two.

The mental files framework can provide another exculpation strategy. Paderewski-Peter became acquainted with *Paderewski* twice: once as a talented musician, and the other time as a politician. Let us assume first that just a single file *a* is opened in Peter's mind to store information gained through this acquaintance relation, roughly the guy playing the piano and the guy making a political speech at a rally. In both cases, it is not the information stored in *a* that constitutes the mode of presentation of *Paderewski* (as in the case of a Fregean Sense or a *de dicto* mode of presentation), but the mental file *a* itself by embodying the acquaintance relations Peter bears to Paderewski. That is to say that the mental file is a de re mode of presentation of Paderewski.

Now, Paderewski-Peter's prejudice against politicians "splits," so to speak, Peter-Paderewski's mind in two, giving birth to two mental files cognitively encapsulated from each other: one of them storing information about the talented pianist, say a_1, and the other about the politician, a_2. Given this, even without holding the propositional attitude 3, Paderewski-Peter's contradictory conclusion 5 is excusable. Paderewski-Peter is not to be blamed for 5. Thus, Boghossian's criticism seems to be ineffective: the violation of IKCCS does not involve irrationality.

6-Ta_1.

7-~Ta_2.

5-Ta_1 & ~Ta_2.

Is this exculpation acceptable? That depends on what we understand by "exculpation." To be sure, Paderewski-Peter is doing his best from his own subjective perspective. From that subjective perspective, he is not to be blamed. However, as Sorensen remarks, "individualists take norms of rationality to be egoistic: only the individual's intellectual self-interest counts" (1998, p. 330). Yet, we should not confine rationality assessments to the individual's narrow psychology as in the cases given by using the mental files framework. We should take a more social stand. So, if "exculpation" is understood as a refusal of the fact that Paderewski-Peter is really committing a contradiction, Paderewski-Peter is not exculpable, and the exculpation strategy is doomed to fail. Considering that the propositions expressed by sentences 6 and 7 are singular or object-involving, by thinking 5, as a matter of fact, Paderewski-Peter holds that:

4-Paderewski is a talented musician & Paderewski is not a talented musician.

The exculpation strategy misses the big picture. As the identifier $a_1 = a_2$ is opaque to Paderewski-Peter, we must assume that Paderewski-Peter is also blind to the reasoning that he is performing. Indeed, Paderewski-Peter is blind to the fact that the patterning of reasoning is the simple rule of conjunction-introduction (from 6 to 5 above: if we have P and we have Q, we must have P&Q).

3 | THE VIOLATION OF TRANSPARENCY OF DIFFERENCE

The details of Twin-Earth are well known and hence I am only going to outline the story briefly. In Putnam's seminal paper,[9] the externalist argues that some individual on Earth refers to *water* whenever he entertains WATER-thoughts or deploys the concept WATER. His replica on Twin-Earth refers to XYZ whenever he entertains XYZ-thoughts or deploys the concept TWATER. Yet, this does not threaten the epistemic authority of the first-person in what Burge has called cogito-like thoughts, that is, second-order contemporaneous thoughts about his/her first-order thoughts. If the Earthian's first-order thought "water quenches thirst" refers to H_2O, his introspective second-order thought "I thought that water quenched thirst" also refers to H_2O, without the need for any

[8]See Boghossian, 2015, p. 14.

[9]See Putnam, 1975.

justification whatsoever, because his introspective second-order thought inherits, so to speak, the reference of his first-order thought by embedding it. This is what Burge famously called "self-verification."

However, things change dramatically in the slow-switching cases later introduced by Burge.[10] Let us suppose that Peter has been moved unknowingly from Earth to Twin-Earth several times. Yet, quick-switching would not be a case in which thoughts switched but the introspection remained the same. The idea is that his comings and goings only switch his thoughts and concepts if he remains long enough on either planet to establish the environmental relations necessary for new thoughts. So, let us suppose that after some time on Twin-Earth, he acquires the twin-earthly concept TWATER and begins to think TWATER-thoughts just as everybody else in his new environment.

The first question that Burge's slow-switching case invites is whether Switched-Peter still possesses his old concept WATER or whether his newly acquired concept TWATER has entirely replaced his old one. Here the philosophical intuitions come apart. According to Burge's cohabitation thesis, Switched-Peter will retain his old concept WATER side by side with his new one TWATER, which makes mental tokens of the same mental type "water" become equivocal, sometimes expressing the concept WATER and sometimes the concept TWATER.[11] According to Recanati's fusion account, both files get merged, and according to Tye's presentist account, the new content *twater* replaces the old one *water*.[12] The problem is that this is something that the subject can never know a priori. Thus, content-externalism in slow-switching cases violates the second part of Boghossian's content-transparency claim, namely:

Transparency of difference: If S's faculty of introspection is working normally, then if S occurrently thinks two thoughts that exercise *different* concepts, he will be able to know that fact introspectively (Boghossian, 2015, 98).[13]

The consequence could not be more disastrous. Switched-Peter is now disposed to fall prey to a simple fallacy of equivocation[14]:

A-Whoever floats on water gets wet.

B-Pavarotti once floated on water.

C-Therefore, Pavarotti once got wet.[15]

Let "FW" stand for floats on water, "WET" stand for being wet, and "a" for Pavarotti, prima facie, Switched-Peter's reasoning instantiates the simple rule of the universal quantifier:

A-(X) [FWx ⊃ WETx].

B-FWa.

C-Therefore, WETa.

The problem is that the semantic values of "W" and "a" are varying. According to Burge's cohabitation thesis, in the first premise the word "water" (W) refers to *twater*, while in the second the same word "water" (W) refers to

[10] See Burge, 1998.

[11] According to Burge's cohabitation thesis, the subject retains both concepts instead of replacing the old one with the new one.

[12] I will come back to both positions in the next sections.

[13] Both claims echo Boghossian's formulation:

> (a) If two of a thinker's token thoughts possess the same content, then the thinker must be able to know a priori that they do; and (b) If two of a thinker's token thoughts possess distinct contents, then the thinker must be able to know a priori that they do. (2011, p.1)

And his original formulation of 1994:

> The thesis of the epistemic transparency of content may be usefully broken up into two parts: (a) If two of a thinker's token thoughts possess the same content, then the thinker must be able to know a priori that they do; and (b) If two of a thinker's token thoughts possess distinct contents, then the thinker must be able to know a priori that they do. Call the first the thesis of the *transparency of sameness* and the second the thesis of the *transparency of difference*. (1994, p. 36)

[14] The converse of "univocation" is equivocation. So, the fallacy in question takes place because the reasoner treats ambiguous expression as if it were univocal.

[15] Boghossian, 2015, p. 100.

water. In other words, while the concept TWATER figures in the first premise, it is the concept WATER that figures in the second. Given this, all the premises are true, even though the Pavarotti argument is clearly invalid, since Switched-Peter falls prey to a fallacy of equivocation. Let "FT" stand for floating on twater and what we have is the following:

A-(X) [FWx ⊃ WETx].

D-FTa.

F-WETa.

What makes this fallacy quite odd and hence intriguing, claims Boghossian, is the fact that the equivocation in question does not occur because "some subjects may be sloppy and careless and simply do not pay enough attention to whether they are being inconsistent".[16] Interestingly, continues Boghossian in his objection, the detection and correction of irrationality would not in principle be accessible a priori to Switched-Peter. Thus, assuming content-externalism, Switched-Peter is deemed irrational because he violated the transparency of difference: he cannot know a priori, purely introspectively, that he is exercising two concepts, WATER and TWATER, rather than one.

To be sure, what makes this fallacy quite interesting is the fact that Switched-Peter is not sloppy and careless and, even more interestingly, he is not in any condition to detect and correct himself a priori, that is, without the help of some investigation of his actual environment. However, I want to suggest that such a conclusion is far from being as scandalous as Boghossian claims if one that does not confine rationality assessments to the individual's narrow psychology and takes a more social stand.

4 | IS THERE A FALLACY OF EQUIVOCATION?

There are, in the literature, several ways of deflecting Boghossian's accusation that Switched-Peter falls prey to a fallacy of equivocation in a simple inference. These are further cases of what I have been calling here the compatibilist exculpation strategy. The first noteworthy way of deflecting Boghossian's accusation of a fallacy of equivocation is to assume that the argument is an enthymeme. Garmendia is a case in point.[17] According to him, Switched-Peter misidentifies *water* as *twater* and this misidentification is an implicit premise in his Pavarotti argument:

F-WATER is TWATER.

A-Whoever floats on twater, gets wet.

B-Pavarotti once floated on water.

C-Therefore, Pavarotti once got wet.

On this reading, the argument is entirely valid; however, it is unsound because the implicit premise D is simply false: *water* is not *twater*.

F-W = T.

A-(X) [FWx ⊃ WETx].

D-FTa = FWa.

E-WETa.

In the same exculpation strategy vein, Recanati distinguishes cases where identity is asserted from cases where it is merely *assumed or presupposed*. Rather than tacitly thinking F Switched-Peter is assuming or presupposing F: "identity is presupposed, but there is an extra premise, namely the presupposition itself. The presupposition is not built into the modes of presentation, yet it is at work nonetheless" (Recanati, 2012, p.135).

[16] Boghossian, 2015, p. 111.

[17] See Garmendia, 2014.

The major problem of this line of argument is the following: as Switched-Peter has not the faintest idea of Twin-Earth, it is hard to swallow the notion that he believes F. With Recanati, I can hold that identity is presupposed, but against him we cannot maintain that such presupposed identity is an extra premise in the argument.

Here we must go back to something we have seen before in the case of Paderewski-Peter. The best way of capturing what is going on in Peter's mind is in the way suggested by Sainsbury and Tye: Switched-Peter behaves "as if" he believes that he has just one concept rather than two. Or in the way Boghossian puts it: Switched-Peter "treats" the concepts WATER and TWATER as if they were the same. My point here is that regardless of whether Switched-Peter holds F or only behaves "as if" he believes that WATER is TWATER, the fallacy of equivocation is still in place, at least if we embrace Burge's cohabitation thesis. If Switched-Peter does not replace his old concept WATER with his new concept TWATER, he still falls prey to the fallacy of equivocation. And the reason is the same as before: the semantic values of "W" and "a" are being determined by the environments. That is why Switched-Peter violates IKCCD: he cannot introspectively detect that he is deploying two concepts in his reasoning, WATER and TWATER, rather than one.

5 | THE ANAPHORIC SOLUTION

However, by far the most ingenious attempt at deflecting Boghossian's accusation of fallacy of equivocation is Burge's anaphoric reading of the Pavarotti argument (Burge, 1998). As the Pavarotti argument takes place on Twin-Earth rather than on Earth, following Burge's cohabitation thesis, we must assume that in the first premise "water" refers either to *water* or *twater*. The interesting point is that it does not really matter: the choice will be a matter of what is anaphoric based upon the first premise of the reasoning.

But let us suppose that now Switched-Peter is thinking of *twater*. Thus, if we take as the first premise statement A, then "water" refers to *water*, and anaphorically in the remaining premises the word "water" must also refer to *water*.

A-Pavarotti once floated on *water*.

B-Whoever floats on *water*, gets wet.

G-Therefore, Pavarotti once got wet.

Yet, given the cohabitation thesis, if Switched-Peter does remember that Pavarotti was floating on *water*, he may well form the belief that whoever floats on water gets wet. Given this, the argument is not only valid, but also sound.

Let us suppose that Switched-Peter remembers that Twin-Pavarotti floated on *twater*. Now, anaphorically in the remaining premises, the same word "water" must also refer to *twater*. To be sure, there is no fallacy of equivocation, because in all the premises and in the conclusion, both the word "water" and the concept WATER have the same semantic value: *twater*. Therefore, the Pavarotti argument is valid, however unsound since H is false: as a matter of fact, Switched-Peter cannot remember that either Pavarotti or Twin-Pavarotti has never floated on *twater* (Lake Taupo):

H-Twin-Pavarotti once floated on *twater*.

I-Whoever floats on *twater*, gets wet.

J-Therefore, Twin-Pavarotti once got wet.

Now, it is quite obvious that if one inference is valid but unsound, while the other is valid and sound, they cannot be one and the same reasoning![18] Burge's suggestion may be right and there is no fallacy of equivocation and content is preserved by anaphora. The point is that his anaphoric proposal, rather than undermining skepticism about rationality under content-externalism, opens the doors for it. We are not only blind to the contents of

[18]This gets even worse when we consider that the premises might be "formed independently and only subsequently put together in conscious thought" (Schroeter, 2007, p. 609)

the figure in our reasoning, but also to the reasoning we are performing. The moral then is that, if "meaning ain't in the head," reasoning is not either.

6 | THE REPLACEMENT OF CONCEPTS

The third noteworthy exculpation strategy to address Boghossian's accusation of fallacy of equivocation (or fallacy of *univocation*) is Tye's rejection of the cohabitation thesis.[19] The assumption is that after a time on Twin-Earth, Switched-Peter's concept WATER is entirely replaced by the new concept TWATER. Thus, in all the premises and conclusion of the Pavarotti argument, "water" refers to *twater* without equivocation. For another reason, we arrive at the same result as Burge:

G-Twin-Pavarotti once floated on *twater*.

H-Whoever floats on *twater*, gets wet.

I-Therefore, Twin-Pavarotti once got wet.

If that is right, when Peter remembers "Pavarotti once floated on water," the semantic values of "Pavarotti" and "water" are determined by Peter's present environment (Twin-Earth) rather than by his past environment (Earth).[20]

G-FTa.

Now, according to this presentist view, the reasoning is again unsound, because in H "water" means *twater* and Switched-Peter has never seen or thought before that Pavarotti once floated on *twater*.

However, again, the exculpation is only acceptable if "exculpation" is understood from a subjective perspective. However, if one does not confine rationality assessments to the individual's narrow psychology, this is unacceptable as exculpation in the sense that it rules out Switched-Peter's fallacy. I no longer buy the Orwellian view that the present determines the past. I reiterate my diagnostic: the individuation of his reasoning is not only up to Switched-Peter, but also to the environment in which he is embedded. As in the slow-switching cases, there are always two or more environments in which the fallacy of equivocation comes to me as the natural and most intuitive diagnostic.

7 | THE FUSION OF FILES

We have critically examined the cohabitation thesis and the replacement thesis. Now the last exculpation strategy is to assume the fusion thesis. If Switched-Peter does not intentionally believe that *twater is water*, but only *behaves as if* he believes that *twater* is *water* or only *treats* twater as water, one could claim that both concepts WATER and TWATER have fused in Switched-Peter's mind.

The best model here is Recanati's mental files framework.[21] On Earth, Peter got acquainted with the *water* of the lakes, oceans, rivers etc. and hence a mental file WATER is opened in his mind to store information about the stuff by exploiting the acquaintance relation he bears to that stuff. Once on Twin-Earth, Switched-Peter got acquainted with other stuff, namely *twater*, which fills the lakes, oceans, rivers etc. and a new mental file TWATER is opened in his mind to store that information by exploiting the acquaintance relation he bears to that stuff. However, as Switched-Peter ignores the slow switch, those files get *merged* into a single file WATER-AND-TWATER. Even though IKCCD is violated, the Pavarotti argument is valid but unsound:

J-Pavarotti once floated on *water and twater*.

K-Whoever floats on *water and twater*, gets wet.

[19] See Tye, 1998. Actually, this is a position Sainsbury and Tye still hold (see Sainsbury & Tye, 2012, pp. 94–95)". In their reply to Boghossian, 2015, however, they assume the thesis of cohabitation for the sake of argument. See Sainsbury & Tye, 2015.

[20] Again, this is the view I have defended in another paper. See also Ludlow, P. 1995, 1999.

[21] See Recanati, 2012. Recanati himself does not clearly endorse this strategy.

L-Therefore, Pavarotti once got wet.

The reasoning is unsound because Pavarotti has never floated on *water and twater*. But the argument seems valid. The question, as before, is: is this exculpation acceptable? And the answer is the same as before: that depends on what "exculpation" means. I am not going to repeat myself again.

8 | INTRINSIC AND EXTRINSIC FALLACIES

Sorensen (1998) was certainly the first to endorse what I have been calling here the "bite the bullet" strategy. Faria (2009) endorses Sorensen's original insights by claiming that are fruitfully illuminated when set against the framework provided by comparison with the prima facie unrelated topic of moral luck (see 2009, p. 186).[22]

Sorensen's main point is his distinction between what he calls "intrinsic" and "extrinsic" equivocation. He claims: "the distinction between intrinsic and extrinsic inconsistency disarms Paul Boghossian's criticism that externalism leads to logical skepticism" (1998, p. 329). In this view, the accidental semantic tourist in the slow-switching scenario is nothing but an exotic extrinsic equivocator.

To begin with, what is a case of intrinsic equivocation? Sorensen provides us with a classical example of the fallacy of equivocation: "the end of a thing is its perfection; death is the end of life; hence death is the perfection of life" (1998, p. 321). To be sure, we learn from the manuals of fundamental logic that to avoid the classic fallacy of equivocation, we should keep separate the different meanings involved; in this case keeping separate the different meaning or concepts of "end.".

What about the case of "extrinsic" equivocation? Sorensen's examples here are what he calls "indexical" words like "water". According to him, the simple recognition of those "indexical" words or concepts is enough to assume that there are extrinsic fallacies of equivocation in contrast to intrinsic ones. He exemplifies with this:

M-Here it is hot.

N-Here it is humid.

O-Therefore, here it is hot and humid.

The point is that even internalists must recognize this kind of extrinsic equivocation. Given this, if extrinsic equivocation is the target of Boghossian's logical skepticism, the target is missed since externalists and internalists are equal victims.

Putnam (1975) was the first to characterize those words as "indexicals." Yet, that was a great mistake. They are not indexicals in any plausible sense. First, their meaning does not remain unchanged when their reference varies from context to context (like indexicals). On the contrary, the meaning of "water" is completely different on Earth and Twin-Earth. Second, indexicals have meaning by courtesy. Their meaning is just a rule that maps context onto contents. In contrast, "water" is not a rule and has a reference fixed once and for ever. Third, indexicals do not have a content of their own, in opposition to "water," but merely a content-schema. Thus, while the indexical adverb "here" has the same content-schema, delivering different places as contents in different contexts, "water" has the same content in all possible worlds: H2O. Fourth, environment here is not a context in the sense of indexicals. Fifth, while my mental states on Earth and the mental states of my twin on Twin-Earth are much the same whenever I think or utter "here" on Earth and he thinks or utters "here" on Twin-Earth, indicating the same place, my mental state on Earth and the mental state of my twin on Twin-Earth are quite different whenever I think or utter "water" on Earth and he thinks or utters "water" on Twin-Earth.

[22] However, he shows reservations about:

> Sorensen's unabashedly consequentialist approach to blame-worthiness; unlike Sorensen, moreover, I will make essential use of a distinction between excusable and inexcusable ignorance: something which the prevailing approach to the 'externalism and inference' debate (relying, as it does, on the 'slow switching' thought experiments introduced by Burge in 1988) has made all but invisible. (2009, pp.186–187)

Now, let us consider Boghossian's transparency claim again:

Introspective knowledge of comparative content: When our faculty of introspection is working normally, we can know a priori via introspection whether any two present, occurrent thoughts exercise the same or different concepts.

By carefully distinguishing proper indexicals from Twin-Earthable words and concepts such as "water," Boghossian could easily concede that his transparency thesis does not extend to indexicals because these do not have contents own their own. Again, indexicals are only content-schema. Thus, even when our faculty of introspection is working normally, we can know a priori via introspection whether two uses of the indexical "here" have the same or different contents because outside the context they do not possess contents in the first place. Therefore, Sorensen's appeal to indexicals cannot disarm Boghossian's criticism that content-externalism leads to logical skepticism (1998, p. 329). Boghossian could insist that, as we cannot detect a priori by introspection alone the difference between water and twater, content-externalism leads to logical skepticism.

Moreover, Sorensen's distinction between "extrinsic and intrinsic" fallacies clearly relies on a local content-externalism in contrast to a global content-externalism. While local externalism claims that several concepts are Twin-Earthable, but certainly not all, global content-externalism does not recognize the distinction between Twin-Earthable and not Twin-Earthable concepts. Now, if we do not recognize the distinction between Twin-Earthable and not Twin-Earthable concepts, the imposing conclusion is that there are no clear boundaries between extrinsic and intrinsic fallacies; after all, all concepts are type-individuated, in part, by intentional and non-intentional relations to the environment and to the local dialect. So, in Sorensen's words: "when there is only a borderline difference, there is no fact of the matter." (1998, p. 324)

Let us return to the classical example from manuals of logic of a fallacy of equivocation: "the end of a thing is its perfection; death is the end of life; hence death is the perfection of life" (1998, p. 321). If we follow Sorensen and describe what we have here, the answer is an *intrinsic* fallacy of equivocation. We are assuming that the reasoner belongs to a single English-speaking community, he possesses the introspective ability to distinguish a priori the different meanings of "end" or concepts END, but hasn't exercised this ability dues to sloppiness, neglect, or inattention.

Now, let us assume that the word "end" or the concept END behave just like the word "water" or the concept WATER. Imagine two English-speaking communities, A and B, without any contact between them. In the first community A, the word "end" means only *aim*, while in the other community B, the word means only conclusion of a task. As usual, let us suppose that we have a semantic traveller going from A to B in a classic slow-switching scenario. Still without assuming the local dialect, our semantic traveller may remember what his old man always said to him:

P-The end (aim) of a thing is its perfection.

Yet, after having assimilated the local dialect of B, he comes to think this:

Q-Death is the end (conclusion) of life.

And he comes to the infelicitous conclusion:

R-Hence death is the perfection of life.

This fallacy of equivocation should be *extrinsic* by Sorensen's account. But the question is: do we have any criteria to separate this infelicitous reasoning from the classic fallacy of equivocation that Sorensen calls the *intrinsic* one? The one criterion that Sorensen could appeal to in order to draw a sharp boundary between intrinsic and extrinsic fallacies of equivocation is exactly what he deplores the most: the assumption that we should confine rational assumption to narrow individual psychology.[23]

Thus, I see little point in claiming that this fallacy of equivocation is *intrinsic or extrinsic*. Once one embraces global content-externalism rather than local content-externalism, there is no fact of matter that could separate intrinsic from extrinsic equivocations. The content-externalist must bite the bullet and swallow it too.

[23]See Sorensen, 1998, p. 330.

9 | CONCLUSION

To start with, Boghossian's Introspective knowledge of comparative content is unreasonable even if we leave content-externalism aside. For one thing, it tacitly assumes something that is hard to swallow, namely that introspective self-knowledge is discriminative like perceptual knowledge. It seems reasonable (even though disputable) to claim that if I perceive, say, a yellow cube straight ahead of me, it is because I am able to discriminate it from other particulars in my visual field and also able to single it out from its background. Yet, introspective self-knowledge is not self-perception of our own concepts and thoughts! Suppose I formed a demonstrative concept of such a yellow cube the first time that I saw it. When I see a qualitatively identical but numerically different yellow cube in exactly the same place, I may form a second demonstrative concept. The problem is: even assuming content-internalism, that is, the claim that content is individuated entirely by what is going on inside my skull (I and my twin on Twin-Earth possess the same concept WATER), can I tell just by introspection one concept from the other? Can I know just by introspection that I am entertaining one or two demonstrative concepts of yellow cubes straight ahead of me? I do not believe so.

As I mentioned in the introduction, rather than looking for a compatibilist exculpation strategy that explains away both equivocation and "univocation" fallacies under content-externalism, I have claimed that one must embrace a global form of content-externalism and, hence, not only bite but also swallow the bullet. If from a psychological or individual perspective Paderewski-Peter's and Switched-Peter's reasoning is not to blame, from a social perspective it turns out to be not only unsound but also invalid. But the moral to be drawn is that the Paderewski-Peter and Switched-Peter cases clearly show that a successful reasoning depends not only on the subject's ability to follow rules of inference. It also crucially depends on something that is completely out of his/her control: the environment. Now, since in an externalist view in general, not only mental states but also abilities are individuated not only by what is going on inside the skull, but also by whatever external, mind-independent particulars (if any) to which the individual relates, what I am defending here is what I am calling biting and swallowing the bullet. For one thing, as the semantic value of our concepts depends on the environment, the reasoning depends on our non-intentional relations to what is going on outside our skulls.

Let us return to the Paderewski-Peter case. He aims to follow a simple conjunction-introduction rule (if P and if Q, then P&Q), but ends up in a contradiction (if P and if ~P, then P&~P). Why is this so? Because his simple reasoning can only be understood in light of his non-intentional relations to his environment. The very same thing can be said about Switched-Peter. He aims to follow the simple rule of instantiation of a universal quantifier but ends up falling prey to a simple fallacy of equivocation. Why is this so? Again, because his simple reasoning can only be understood in light of his non-intentional relations to his environment. In both cases, I claim that the non-intentional relations between the reasoner and his environment individuate the reasoning itself. I conclude that if reasoning is "formally" individuated by the logical form that it exemplifies, it is also "materially" individuated by intentional and non-intentional relations that the reasoner bears to his/her environment and to his/her speech community.

ORCID

Roberto Horácio Sá Pereira https://orcid.org/0000-0002-9117-0755

REFERENCES

Boghossian, P. (1992). Externalism and inference. *Philosophical Issues*, *2*, 11–28. https://doi.org/10.2307/1522852

Boghossian, P. (1994). The transparency of mental content. *Philosophical Perspectives*, *8*, 33–50. https://doi.org/10.2307/2214162

Boghossian, P. (2011). The transparency of mental content revisited. *Philosophical Studies*, *155*(3), 457–465. https://doi.org/10.1007/s11098-010-9611-3

Boghossian, P. (2012). What is inference? *Philosophical Studies*, *169*(1), 1–18. https://doi.org/10.1007/s11098-012-9903-x
Boghossian, P. (2015). Further Thoughts on the Transparency of mental content. In S. C. Goldberg (Ed.), *Externalism, self-knowledge and skepticism, new essays* (pp. 95–97). Cambridge: Cambridge University Press.
Burge, T. (1998). Memory and self-knowledge. In P. Ludlow & N. Martin (Eds.), *Externalism and self-knowledge* (pp. 111–127). CSLI.
Faria, P. (2009). Unsafe reasoning: A survey. *Dois Pontos*, *6*(2), 185–220.
Faria, P. (2010). Memory as acquaintance with the past: Some Lessons from Russell, 1912–1914. *Kriterion: Revista de Filosofia*, *51*(121), 149–172.
Faria, P. (2015). Inferential rationality and internalist scarecrows. *Manuscrito*, *38*(3), 5–14.
Garmendia, M. E. (2014). Externalism, rational explanation, identity premises. *Teorema*, *XXXIII/3*, 31–48.
Kripke, S. A. (1979). A puzzle about belief. In A. Margalit (Ed.), *Meaning and use* (pp. 239–283). Dordrecht, The Netherlands: Reidel.
Ludlow, P. (1995). Social externalism, self-knowledge, and memory. *Analysis*, *55*, 157–159.
Ludlow, P. (1999). First person authority and memory. In M. De Caro (Ed.), *Interpretation and causes: New perspectives on Donald Davidson's philosophy* (pp. 159–170). Dordrecht, The Netherlands: Kluwer.
McKinsey, M. (1991). Anti-individualism and privileged access. *Analysis*, *51*(1), 9–16.
McKinsey, M. (2002). Forms of externalism and privileged access. *Philosophical Perspectives*, *16*(s16), 199–224.
McKinsey, M. (2018). Skepticism and content externalism. *Stanford Encyclopedia of Philosophy*.
Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, *7*, 131–193.
Recanati, F. (2012). *Mental files*. Oxford, UK: Oxford University Press.
Sainsbury, R. M., & Tye, M. (2012). *Seven puzzles of thought*. New York: Oxford University Press.
Sainsbury, R. M., & Tye, M. (2015). Counting concepts: Responses to Paul Boghossian. In S. C. Goldberg (Ed.), *Externalism, Self-knowledge and skepticism, new essays* (pp. 113–119). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107478152.
Schroeter, L. (2007). Illusion of transparency. *Australasian Journal of Philosophy*, *85*(4), 597–618. https://doi.org/10.1080/00048400701654820
Sorensen, R. A. (1998). Logical luck. *The Philosophical Quarterly*, *48*(192), 319–334. https://doi.org/10.1111/1467-9213.00103
Tye, M. (1998). Externalism and memory. *Aristotelian Society Supplementary*, *72*(1), 77–94.

How to cite this article: de Sá Pereira RH. Dangerous Liaisons. *Ratio*. 2019;00:1–13. https://doi.org/10.1111/rati.12241