UNIVERSITAT DE BARCELONA

Thinking Alike:
Five Essays on the Publicity of Thought

Pensar el mateix:
cinc assaigs sobre la "publicitat" del pensament

Matheus Valente

UNIVERSITAT DE BARCELONA

PHD THESIS

Thinking Alike: Five Essays on the Publicity of Thought

Matheus Valente

Barcelona
2020

THINKING ALIKE: FIVE ESSAYS ON THE PUBLICITY OF THOUGHT

Matheus Valente

Pensar el mateix: cinc assaigs sobre la "publicitat" del pensament

Doctoral thesis submitted by Matheus Valente to the University of Barcelona for the degree of Doctor of Philosophy
Supervisor: Manuel García-Carpintero
Co-supervisor: Josep Macià

University of Barcelona
Faculty of Philosophy
PhD Program: Ciència Cognitiva i Llenguatge/Cognitive Science and Language
(EEES: HDK08 Ciència Cognitiva i Llenguatge)
Research Line: Philosophy of Language, of Mind and of Science

June 2020

CONTENTS

INTRODUCTION

THE PAPERS

INTRODUCTION

1 An interconnected whole

October 2016 feels like a long time ago. Back then, just about to start my PhD in Barcelona, I had little idea of what my dissertation was going to be about. Unlike many of my colleagues, I had no overarching hypothesis over which to dwell for the three or four years to come. It surprised me that many of them not only already knew what they wanted to do research on but could delineate their main conjectures with enviable precision.

My move from Brazil to Barcelona around that time was made possible by being one of the 14 fortunate recipients of an incredible scholarship, part of the bigger-than-life Diaphora project[1], spearheaded by Sven Rosenkranz and countless other philosophy luminaries across Europe; the bigger the prize, the greater the responsibilities – still, I couldn't but stutter when people, surely trying to be nice to the newcomer, asked "what's your thesis about?".

In retrospect, all I had back in October 2016 was a vague impression: there is little consensus on what —if anything— is for two persons, or for the same person at different times, to think the same thoughts or to entertain the same concepts. These two interrelated issues are sometimes called, respectively, the problem of thought publicity and of cognitive dynamics. The second issue was to finally lose protagonism in my dissertation, being progressively relegated to the background; thought's publicity, on the other hand, consistently appeared as a central aspect of the papers I read, and, unsurprisingly, in the papers I wrote. It is thus no surprise that the dissertation is called what it is.

I confess I was slightly relieved when, just very recently, I looked back on the papers I had written during my doctorate years and discovered that, indeed, they do form a – sort of – interconnected whole. I now explain why this is so. I'll do this in two steps. First (section 2), I attempt to briefly outline the methodological stance which informs most of the chapters that follow. I do this by discussing some fundamental recurring themes, as well as fending off a few preliminary criticisms that would make my project

[1] The Marie Skłodowska-Curie European Training Network DIAPHORA served as a European research and training platform for collaborative research on the nature of philosophical problems, their resilience and the sources of persistent divergence of expert opinion about them, and their relation to conflicts in the practical sphere. It received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 675415. For more: http://www.ub.edu/diaphora/

seem unfairly unattractive. The starting point is that thoughts and concepts are functional entities, and thus definable by their constitutive theoretical roles. Which roles these are, of course, is one of the most pressing issues examined, but the focus is on a particular one: thought's role in explaining several intersubjective relationships such as communication, understanding, and genuine disagreement. Secondly (section 3), I tell you a little bit about each chapter and explain how they swirl around these topics, regardless of the occasional terminological discrepancy or divergent focus. In any case, don't worry: I will avoid detailing everything that the papers talk about and mainly focus on their mutual interactions.

2 Recurring Themes

As any old theoretical entities, thoughts (the whole thing) and concepts (its parts) are supposed to earn their keep by means of the jobs they can do – where 'doing', in this case, consists in being able to adequately explain or at least illuminate phenomena which can itself be described independently of them. Which phenomena are thoughts and concepts supposed to explain? Crucial to this dissertation is Publicity – thought's role in an account of several intersubjective relationships between thinkers, especially successful communication, understanding and genuine agreement.

Four of the five papers compiled therein have Publicity at their forefront, while the exception (chapter 4) discusses issues very much in its vicinity. But Publicity is a tough role to accommodate. On each of these four chapters, I pair Publicity with another important thesis - either a distinct role that thoughts are equally expected to play, or a more precise view about concepts – and the ensuing result is inevitably the same: inconsistency. Thus, in Chapter 1, we see that Publicity clashes with *Explanation*, the theoretical role according to which sameness of thought allow us to predict sameness of behavior. Chapter 2 is about Publicity's conflict with *Frege's Constraint*, the principle that thoughts must track cognitive or informational import. Chapter 3 and Chapter 5 put Publicity face to face with, respectively, *internalist* and *relationist* view of concepts. Unsurprisingly, the predicament is the same: either Publicity or the accompanying view must be rejected.

In summary, all but one of the papers compiled herein can be advertised as a conflict between Publicity and another contending claim. Given that each of those

contenders have independent plausibility behind them, the fact that all of them seem to clash with Publicity amounts to an astounding cumulative argument against it. It would be nice if it could be otherwise, but the way things have gone, we cannot but conclude: Publicity must go.

In the next couple of subsections (2.1, 2.2), I hope to explain why this conclusion is not to be taken too lightly. Publicity, I claim, is a highly desirable principle for more than one methodological reason. If it must really go, this is no reason to celebrate. Besides, rejecting it is much trickier than has been thought. In subsection 2.3, I fend off a few *prima facie* criticisms against it that sound intuitive at first look, but ultimately fail. This will help us set the stage for what I take to be the appropriate reasons for letting Publicity go, which will be spelled out in the following chapters.

2.1 Thoughts and Publicity

'Thought' is, at least in restricted analytic philosophy circles, a term that evokes the philosophical work of Frege. In *On Sense and Reference*, Frege explains what he means by 'thought' in a discreet footnote: "by a thought I understand not the subjective performance of thinking but its objective content, which is capable of being the common property of several thinkers" (Frege, 1892, fn. 5).

This telling footnote has several aspects, but a clear thread underlies it: thoughts are not subjective (such as my current thinking about lunch right now, with all the contingent associative ideas that surround it), but objective, and thus shareable by distinct thinkers. Armed with an opportunity to explain what thoughts are, Frege thus basically sets out the basic ideas behind what me and other authors call 'Publicity'.

Frege's suggestion takes us far, but not far enough. It is ok to suggest that thoughts are introduced in order to explain objective, and not subjective, aspects of thinkers – but which aspects? Here we might start to deviate from Frege, but the cornerstone idea of this dissertation, put in the most general and pre-theoretical terms as possible, is that the notion of thought must help us explain when two thinkers think alike – in other words, when two thinkers are thinking the same. This, then, is the phenomenon we can describe prior to the introduction of thought *qua* theoretical notion: subjects sometimes think alike – e.g. when one takes the other at her word and believes what she says – and sometimes do not – e.g. when they're unknowingly talking past each other.

The promising idea then is: by introducing thoughts to our theoretical framework, we can systematize that pre-theoretical phenomenon and illuminate its inner workings.

We can thus define Publicity as the claim that thoughts (or at least the relation of sameness of thought)[2] must help us account, in a reasonably direct fashion, for several intersubjective relationships between distinct thinkers, such as successful communication and understanding. These intersubjective relationships are supposed to be a first-pass precisification of the admittedly vague idea that sometimes subjects think alike and sometimes not. The intuitive connection between the latter idea and the intersubjective relationships is then naturally hypothesized to work in the following manner: if two subjects successfully communicate and understand each other, then (*modulo* cases of insincere utterances or suspicious hearers, which I'm sorry to say I won't be concerned with in this dissertation) they think alike.

Let me emphasize how helpful this precisification is supposed to be. Our commonsensical intuitions about when people think alike with respect to a certain subject matter are not always crystal clear; but they're certainly sharper with regards to whether two subjects have successfully communicated or understood each other. Take, for example, a layperson who knows no more about megabytes than that they're stuff you can measure "the size" of computer programs with, and a computer specialist who could give you an accurate description of how they function. Do these people think the same as each other if, for example, both assent to 'floppy disks can store 1.44 megabytes'? Well, in a sense there is no principled way to answer that question – "think the same in which respect?", one might wonder. When the layperson assents to that, she's mainly focused on e.g. a computer game whose size is just a bit under 1.44mb, so she could carry it in one floppy. When the specialist assents to that, she might be focused on… I don't even know, complicated stuff about bytes and bits, I guess.

On the other hand, if we rephrase the question as "would the layperson understand the computer specialist if the latter had uttered that sentence (or vice-versa)?" then our verdicts might point with a bit more strength towards a positive answer. It is easy to conceive of communicate scenarios in which one utters that sentence to the other, and both leave knowing more than they did before. Some could, of course, still raise their

[2] I will usually speak of *thoughts* as the key theoretical entities, but it might very well be that we just need an equivalence relation, *sameness of thought*, in order to do all we need to do.

eyebrows and wonder whether this instance of communication was really 100% successful, or whether understanding there was not only "partial". I welcome that kind of hypothesis. However, I find it hard to deny that this question, unlike its analogue from the previous paragraph, has already taken us to the field of philosophical doubt. It is not the question the ordinary person in the street would ask, but one which is raised inside philosophy departments by people who wonder about the roles the concepts of successful communication and understanding should play. The distinction between the philosopher and the layperson is, I admit, not clear-cut. But limits need to be traced at some place in order for inquiry to get off the ground; this is where I've traced them.

Of course, intuitions about cases are important but not irrefutably dogmatic. As Austin nicely puts in a discussion about the use of ordinary language within philosophy:

> Certainly, then, ordinary language is not the last word: in principle it can everywhere be supplemented and improved upon and superseded. Only remember, *it is the first word*. (Austin 1979, p. 185, emphasis mine)

We can then summarize my most general methodological stance as an inference that starts with a yearn to provide an account of what it is for two subjects to think alike with respect to a certain subject matter, precisifies that notion as consisting in the intersubjective relationships previously mentioned, and ends with the introduction of thought (or sameness of thought) as the minimal notion that would be enough to fulfill that task in a suitably systematic fashion. The question we started this section with was: what are thoughts? I can now give you a minimal answer that mirrors some of what Frege said in that footnote: thoughts are the theoretical posits whose sharing by subjects explain why they've successfully communicated or understood each other. This general idea and understanding of what thoughts are I refer to as 'Publicity' in chapters 3 and 5, as 'Agreement' in chapter 1, and as 'the thought-transfer model of communication' in chapter 2.[3] These distinct names reflect the distinct commitments and objectives of each separate chapter, but their common core all point to the same direction.

Let me pause here and make a disclaimer. I've been implying that my conception of what thoughts are is broadly Fregean, and in sync with what this author says on his

[3] A related but not-quite-the-same principle is called 'C2' in chapter 4.

well-known 1892 paper. This is misleading. For one thing, I have thus far acted as if Frege's quotation never included that bit about "thoughts being contents". If thoughts are contents, then they are semantic notions and prone to have all sorts of properties that semantic entities usually have. At first sight, this doesn't seem to have much to do with Publicity, which is itself neutral on whether thoughts strictly belong to semantics or not (it is not, after all, contradictory to think that whether communication is successful or not crucially depends on non-semantic issues).[4] I might have also led you to believe that Frege took thoughts and intersubjective relationships to stand in a very close, almost indissociable relationship on that paper. But those who have read *On Sense and Reference* might notice that communicative scenarios do not play a very central role in that paper.[5] Instead, Frege's main aim there is to explain how some identities can be informative, and others not, regardless of being made true by the same object's self-identity. This, and not anything I talked about, is the theoretical role Frege mostly cared about back in 1892.

I never meant to do any accurate exegesis of Frege's work, but some clarifications are in order. In providing them, I also intend to make it clearer how Publicity connects with the Fregean discussions about the informativity of identities and other themes. Let us then make a brief detour.

2.2 Frege's Constraint

Frege's main aim in his famous paper, solving the puzzle of informative identities and showing how propositional attitudes with the same truth-conditions can be cognitively distinct, also plays a central part in this dissertation. Indeed, this theoretical role of thoughts divides the spotlight with Publicity in chapter 2 of this dissertation, and appears in the background of all of the others. Throughout the chapters to come, it is variously called 'Frege's Constraint' (chapters 2, and 3), CS-Role (for 'cognitive significance', in chapter 5), and referred to *en passant* as 'the hyper-intensionality of singular thought' in chapter 1. I take Frege's Costraint (as we might choose to call it, following the likes of

[4] Indeed, I am, for the most part, neutral about whether thoughts and concepts are to be conceived as semantic entities or not. The only exceptions is chapter 3, where I'm explicitly examining semantic internalist views, where the semantic nature of thoughts and concepts follows as a corollary. In all of the other chapters, on the other hand, thoughts and concepts are whatever play a certain set of theoretical roles, and these roles will consistently be neutral on that respect.

[5] Although they of course do in his 1918 paper, *The Thought*.

Schiffer 1978 and Recanati 2012) so seriously that I almost never even conceive of rejecting it. Indeed, the only place where I discuss its rejection is within a brief discussion on section 6 of Chapter 2 – but I quickly dismiss that possibility given residual complications that would arise.

My dissatisfaction with Frege's Constraint is not that I do not find it fundamental enough. It's that, if this is the only role for thoughts we have to go on, then their theoretical usefulness will be limited to particular time-slices of single individuals. In other words, Frege's Constraint only enables us to individuate the thoughts that a single subject has at one and the same time. I won't go over the (boringly familiar) details about why this is so here; this will be clear enough in the chapters to follow. Suffice to notice that Frege's Constraint is a principle about what a (single) rational thinker can believe simultaneously, and thus it is useless if we're interested in intersubjective and/or diachronic scenarios.

This then is another substantial reason in favor of the methodological stance I have outlined in the previous section: subscribing to Publicity, and thus ensuring that sameness of thought walks hand in hand with the intersubjective relationships, enables us to have faith that an intersubjective notion of thought, equally capable of characterizing single individuals' cognitive states as of distinct ones, can be designed.

I hope this is seductive enough to convince the reader that Publicity is a highly attractive principle. If not, I presume you will find my attempts to motivate, defend, and ultimately show the limitations of Publicity in the following chapters a bit similar to Quixote fighting windmills. I am however confident that my discussions will have something for everybody, even for those who are suspicious of the idea that communication and understanding require the transmission of a thing (or some kind of sharing), where this thing is also supposed to be what characterizes the cognitive value of our beliefs. Even if you sympathize with this suspicion, you might still appreciate to learn that some reasons to be skeptical of Publicity and my whole methodological stance are better than others. If all else fails, I might at least rest content with having separated the wheat from the chaff, that is, good reasons to be skeptical of Publicity from bad ones – and there's a lot of chaff.

2.3 Much chaff for just a few grains of wheat

Here's a first criticism I'd like to pre-empt: "why should I care if thought can be made intersubjective or not? As long as I can distinguish between Lois Lane's beliefs about Clark Kent from her beliefs about Superman, why would I need anything more? Aren't you sweating for nothing?"

My first reaction to that criticism is that any conception of propositional attitudes according to which they're not a common currency between distinct individuals would fail to give us one of the central constituents of folk-psychology, which is expected to be reflected on any serious cognitive psychology: generalizable intentional explanations. I do not need to go into details here, but the idea is old and familiar: if a subject intentionally acts some way because of the beliefs and desires she holds, then, if another subject also holds those same beliefs and desires, she will, all things being equal, act in the same way. Behavior can not only be explained on the basis of a subject's attitudes, we can also predict that people who believe alike will behave alike. This fundamental principle, which is referred to as 'Explanation' in Chapter 1 (where it is presented as Publicity's contender) and discussed under the rubric of 'Intentional Explanation' in Chapter 5, depends on the idea that distinct subjects can have the same propositional attitudes; from there, it is a quick step to the conclusion that Explanation depends, at least weakly, on the shareability of thoughts.

Leaving these issues to the side, another skeptic might put forward the following complaint: "explain to me again why we should want an intersubjective notion of thought which is as fine-grained as the intrapersonal one carved up by Frege's Constraint? For one thing, Frege's reasons for introducing thoughts in *On Sense and Reference* seems based on an intrapersonal puzzle that doesn't obviously have an intersubjective analogue. Which intersubjective puzzle(s) would this conception of thought be supposed to explain?"

This challenge is well received but, I think, can be met. For each of the intersubjective relationships we've discussed above, we can construct a puzzle which is analogous to Frege's puzzle of informative identities in a crucial respect, namely, in showing that sameness of truth-conditions (or referential content) is not sufficient to account for some target notion. Frege's puzzle of identities shows it is not sufficient to characterize the cognitive profile of our attitudes (since token attitudes with the same

truth-conditions can be cognitively distinct). Analogous puzzles show it is not sufficient to characterize the success conditions of the intersubjective relationships. Communication, for example, can fail even if the hearer acquires, as a result of interpreting the speaker's utterance, a belief which has the same truth-conditions as the speaker's. I refer to these cases as 'Loar-cases', given their inspiration on a thought-experiment found in Loar (1976). Loar-cases are the centerpiece of chapter 4 (indeed, this chapter contains no fewer than 3 variations of these cases), but they also pop up in a couple of other places throughout the dissertation.

To be sure, Loar-cases only show that sameness of truth-conditions is not sufficient for successful communication or understanding, but nothing about them forces us to conclude what sameness of thought then should be. Taking sameness of thought as the key explanatory relation in these cases is to be seen as an inference to the best explanation, and not as deductive consequence of the data. Whether it really is the *best* explanation or not is – of course - one of the main questions I examine. There surely is a lot of good reason for skepticism, I admit. But I repeat: there are many bad reasons around being taken for good ones, and very few good ones that are being properly focused on.

I thus insist on discussing a set of bad reasons to be skeptic about Publicity before proceeding to give you a summary of the papers and finishing this introduction. The first are related to indexicals, the second, to conceptual variability.

2.3.1 Publicity and indexicals

Back in 2017, during my first ever PhD talk at the University of Łódź, I received a question during the Q&A session that haunted me for quite some time. This episode centered around the following accusation: "why would anybody think that communication requires sameness of thought between speaker and hearer? Just look at any instance of communication involving indexical expressions – these are clear cases where successful understanding requires the hearer to entertain a thought which is radically distinct from the speaker's, e.g. if you say 'I', then I must say 'you'!"

On those days, I still hadn't thought long about the subject, and the confident smile of my interlocutor did not help me get ahold of myself and construct a proper reply. I can do better now. I think this type of argument ("indexicals show communication often requires distinct thoughts") is much harder to sustain than people usually presume. It

rests on an unspoken assumption which sounds natural but is actually dubious. The assumption is roughly: no belief expressible by an indexical i^1 is characterizable by the same thought as another belief expressible by an indexical i^2, where i^1 is different from i^2. Let us call this thesis 'Distinctness'. For the sake of exposition, we might limit our discussion to the particular case of I/you. The restricted Distinctness then comes out as: no *de te* belief (expressible by the second person pronoun) is ever *the same as* (characterizable by the same thought as) a *de se* belief (expressible by the first person pronoun).

People often assume Distinctness implicitly, and then jump to the conclusion that Publicity is incompatible with indexical communication. But what are the arguments in its favor in the first place? This is not always easy to answer. I've come up with three different ones: (i) the argument from revised Frege's Constraint, (ii) the argument from the Lewisian view of propositional attitudes, (iii) the argument from Explanation.[6]

Let me now explain why I think (i) fails and why (ii) is not convincing. I hope you'll agree that these discussions are worth the ride. As for (iii), my final verdict is that it is plausible. However, this argument will not be discussed here but in Chapter 1. Indeed, a great part of Chapter 1 is devoted to developing that argument and shielding it from objections. Its plausibility is thus not straightforward, it must be earned. In any case, the argument is not a direct refutation of Publicity. It at most allows us to conclude that Publicity is in conflict with Explanation with regards to indexical scenarios. One could save Publicity by either rejecting Explanation or restricting it to non-indexical thoughts. In summary, I see only one promising argument in favor of the Distinctness and it is neither easy to motivate nor a direct refutation of Publicity. My interlocutor was surely taking too much for granted.

2.3.1.1 Publicity and the revised Frege's Constraint

Frege's Constraint, as we have seen, entails that, if an individual can simultaneously believe of the same object that it has and doesn't have a certain property

[6] Another possible argument is "the argument from immunity to error through misidentification". It would be based on the fact that *de se* beliefs are (sometimes?) IEM, *de te* ones, not. It is not common to employ IEM in the service of individuating thoughts, but this could be the basis for an interesting argument. In any case, I leave it unexplored for the time being.

(simultaneously), then his belief and disbelief should be characterized by distinct thoughts. Can Frege's Constraint be used to argue for Distinctness?

One might try to do so by arguing that there is no pair of *de se* and *de te* thoughts such that it would be irrational to believe one while disbelieving the other simultaneously. Take, for example, the belief Newman expresses by 'These pretzels are making me thirsty!'. It is plain to see that he can have that belief without believing what he'd normally express by the corresponding second-person utterance of 'These pretzels are making *you* thirsty!' – and that is true even if that utterance were directed at Newman himself (e.g. if he were looking at himself in the mirror without realizing it was his own reflection). From this, one might try to infer Distinctness, i.e. that every *de se* and *de te* beliefs are such that they need to be characterized by distinct thoughts.

This might be sound reasoning but it's not enough to support Distinctness. The cases which are of interest to our present discussion are ones where the pair of *de se* and *de te* thoughts are not tokened by one and the same thinker, but by distinct participants in a conversation - a speaker and a hearer.[7] We are, after all, trying to see if indexical communication is compatible with Publicity. This requires a scenario with two thoughts, where one is the result of interpreting an utterance which expresses the other. The two thoughts must be connected by means of an interpretation relation, so to say. More particularly, the kind of case we are interested is one where the speaker (e.g. Newman) makes a first-person utterance, and where the hearer (e.g. Elaine) forms, as a result of interpreting his utterance, a belief she'd express by means of a second-person utterance:

Newman: these pretzels are making me thirsty!
Elaine: these pretzels are making you [addressing Newman] thirsty!

At first sight, Frege's Constraint is completely silent about interpersonal cases such as the one above. It only outputs predictions for thoughts that one and the same subject can hold at one and the same time. Can we somehow expand it so that it helps us account for conversational contexts such as Newman & Elaine's? One way we could try to do this is by asking ourselves whether it would be rational for someone to accept

[7] Strictly speaking, the speaker and the hearer can be the same person as long as she is interpreting her own utterance from the third-person.

Newman's 1st-person utterance as expressing a true thought while taking Elaine's 2nd-person one to express a false one. But the answer to that question is clearly 'yes'. If one mistakenly thinks that Elaine's utterance is directed at a third participant of the conversation, say George, instead of Newman, one might think that Newman's utterance says something true, but that Elaine's doesn't. Just to be clear, if one commits that mistake, one would end up with a bad interpretation of Elaine's utterance, but that doesn't mean this person would be irrational. In any case, it seems hasty to jump from that possibility to the conclusion that Newman and Elaine's utterances express distinct thoughts. Instead, the possibility of taking contrasting attitudes to the thoughts expressed by these utterances seems best accounted by the fact that this mistaken thinker fails to understand who Elaine is talking to in the first place. The case here is similar to one where a subject mistakenly believes that 'vixen' and 'female fox' mean different things, and thus is able to rationally take contrasting attitudes to thoughts expressed by utterances which only differ in the substitution of one term for the other.

These comments suggest that we should focus only on thinkers who successfully understand who the speaker is talking about. Frege's Constraint can then be revised in the form of the following test: two utterances express distinct thoughts if and only if a subject who understands who these utterances are directed at could rationally take contrasting attitudes to them (believing one while disbelieving the other).

Let us then conceive of a thinker who understands that Newman's utterance expresses a *de se* thought about Newman and that Elaine's utterance expresses a *de te* thought about Newman. Can she take contrasting attitudes towards them? It seems she cannot. If that thinker understands that the two utterances are directed at the same person, then there seems to be no rational basis for thinking they diverge in truth-value. She believes what one says if and only if she believes what the other does. Thus, if this version of Frege's Constraint is to be taken as a serious indication of thought distinctness, it at least fails to entail that they're different (if passing the test is taken as a necessary condition for thought distinctness) and possibly shows that they're the same (if not passing the test is taken as a sufficient condition for thought identity).

2.3.1.2 Publicity and the Lewisian view

Other arguments for Distinctness are less general and, instead, follow collaterally from theoretical choices made with other objectives in mind. Thus, for example, if one subscribes to the popular Lewisian view of propositional attitudes, according to which beliefs is modeled as the self-ascription of a property, then, as many authors have pointed out, Distinctness follows as a collateral consequence and Publicity crumbles. While it is true that Lewis' (1979) framework has Distinctness as a consequence, I think this is a "theoretical epiphenomenon" of his view, and not one of its essential aspects. Lewis' objectives had very little to do with Distinctness, and thus tweaking his view so as to make it compatible with the former thesis should not affect its most important virtues. What follows are mere sketches of how this proposal could be developed, but hopefully they will be suggestive enough that the reader will see what I'm hinting at.

Lewis' account of the *de se* in terms of the self-ascription of properties has one main objective: accounting for the cognitive value of *de se* beliefs. Lewis wants to explain how it is possible e.g. for Zeus to believe what he'd express by 'Zeus throws down bolts of lightning' without believing what he'd express by 'I throw down bolts of lightning'.

Lewis' way of accounting for his primary objective is making belief very fine-grained: to believe is to self-ascribe a property. The self-ascription part has one immediate consequence: we do not literally have the same beliefs, since each of us is self-ascribing properties to themselves. But this is not an insurmountable problem, since we can at least have beliefs whose content is the same property.

Surprisingly, even if we attain ourselves to the properties self-ascribed according to Lewis' view, Distinctness still follows. Say that a *de se* belief and a *de te* belief *agree* with each other when they're witnesses to an instance of successful communication (e.g. my 'I' and your 'you' when we understand each other). The problem is that, while every *de se* belief will be modeled as the self-ascription of a property F, every corresponding *de te* belief in agreement with it will be modeled as the self-ascription of the property of inhabiting a world where one's addressee is F. In Lewis' view, no *de se* and *de te* belief ever self-ascribe the same property. Thus, Distinctness is true on that view.

But does it have to be so? Is the entailment of Distinctness one of the things that Lewis needed/strived to have? I don't think so. Notice that Lewis' objective has *prima facie* nothing to do with it. It is not one of his main worries, for example, whether Zeus'

de se belief ('I throw down bolts of lightning') is identical to Thor's *de te* belief which he'd express as 'You [to Zeus] throw down bolts of lightning'.

With that in mind, here's two ways in which one might modify Lewis' framework so that (i) it still accounts for the cognitive value of *de se* thought and (ii) it does not entail Distinctness:

1. Model belief as the *group-ascription of a property*. This has been done already (e.g. work on multicentered propositions). On this view, a *de se* and a *de te* belief could be said to group-ascribe the same property: the property of being members of a group where one of the members is F. But there's a potential problem here. In views that employ multicentered propositions, much care is put into making sure that the members of the corresponding group be ordered. This needs to be done because often we are not simply ascribing a property to some member of the group, but to a particular one (the speaker; the hearer etc.). Thus, what these authors really end up with is something like: *the group is x, y (where x = Newman and y = Elaine), and the property being group-ascribed is that we are members of a group where x is thirsty*. Would Distinctness then follow from the fact that Arlette and Brigitte occupy different places in the ordered set modeling the group? I don't think so, since at least the property being group-ascribed would be the same.

2. Model belief as the x-ascription of a property, where x can take values different from 'self', such as 'addressee'. The idea is that some beliefs are modeled as self-ascriptions of properties but others as the addressee-ascription of (the same) property. We can also perhaps group-ascribe, third-personally-ascribe, world-ascribe a property, Newman-ascribe etc. But I do not care about these possibilities right now. As long as we have a *de se* form of ascription alongside a *de te* form, we can say that a *de se* and a *de te* belief are ascriptions of the very same property, but where the first is a self-ascription, the second, an addressee-ascription. This kind of framework has been, so far, not developed (not even by me, sorry!). I would appreciate having a view where *de se* belief would be modeled as self-ascription of a property, *de te*, as the addressee-ascription of a property, *de dicto* belief as the world-ascription of a property and *de re* as the X-ascription of a property (where X is the object the belief is about). Thus, the same property, *being thirsty*, could

be self-ascribed, addressee-ascribed, world-ascribed or X-ascribed. There are many issues that such a framework would raise, but I will leave at that for now.

2.3.2 Publicity and "conceptual" variability

Here's another complaint I've heard a few times: "Publicity is no good; just look at any psychology manual devoted to concepts: 99% of the authors there agree that people's concepts are highly variable, differing from one tiny contextual change to the other; if communication really required people to have the exact same concepts, then successful communication and understanding would be a faraway utopia".

Here's one example of the type of cases this person would have in mind: a recent experiment found that half of native Dutch speakers tested took the arm to end at the wrist ('arm' thus means the limb from shoulder to wrist), while the others took it to end at the hand's fingertips ('arm' thus means the limb from shoulder to fingertips). If half of the people in the Netherlands have a different concept for such an ordinary thing as an arm, then it is surely wishful thinking to think people generally share their concepts.[8]

The unspoken assumption in that argument is, of course, that our concepts and thoughts must be individuated by our beliefs of what some representations apply to, or at least our dispositions to employ them when engaging in certain higher-order cognitive tasks, such as categorization of objects. There is only one chapter in this dissertation which takes this "internalist" assumption on-board: Chapter 3. In that chapter I explicitly discuss what to do with Publicity given so much variability. Indeed, my reply in that paper is sort of pessimistic: Publicity has to be severely weakened and we have to accept that communication and disagreement can proceed in absence of sameness of "thoughts" and "concepts", conceived of in this way.

On the other hand, one has to remember that, except when I'm explicitly engaging with a particular theory of thought, what I refer to by 'thoughts' and 'concepts' might have very little to do with our categorization skills, or even our dispositions to apply words to certain scenarios. By default, thoughts and concepts are posits designed to play certain roles, such as Publicity, Frege's Constraint or Explanation. This might have

[8] This case is reported by Majid (2010); I learned of it through Pagin (2020). The other papers in that volume – *Shifting Concepts* (2020), edited by Teresa Marques and Åsa Wikforss – present a ton of evidence in favor of conceptual variability and include many philosophically instigating dicussions.

something to do with the psychologists' notion of 'concept', or not. In any case, this is not the fight I volunteered to participate in.[9]

3 An overview of things to come

As I've said before, four of the five chapters of this dissertation can be seen as disputes between Publicity and a rival contender. In order of rounds, the contenders are: Explanation (1), Frege's Constraint (2), Semantic Internalism (3) and Relationism (5). Chapter 4 escapes that aggressive narrative with its focus on Loar-cases, but it also contains its fair share of punches and kicks – the fight in that case is between distinct criteria for successful communication.

Chapter 1's main act is a dispute between Publicity and Explanation. More precisely, I show that scenarios involving successful communication and understanding with indexical expressions are peculiar in that they force us to decide between one principle or the other. I laud this as *the* special feature of indexical attitudes – perhaps a good candidate for what is really special about them in opposition to "boring" *de dicto* and *de re* attitudes. A crucial part of this paper is a defense of my main conclusion – that there is indeed a conflict between Publicity and Explanation with regards to indexical attitudes – against Cappelen & Dever's (2013) elusive challenges, especially their action inventory model argument.

Chapter 2's headline event is the puzzle arising from a conflict between Publicity and Frege's Constraint. The puzzle is based on a familiar Kripkean case involving a subject which takes a thing to be two, and takes its name also to be two. Since she does that, she ends up having two thoughts where a "normal" person has only one. Considerations about communicative success, I argue, allow us to conclude that Frege's Constraint – the principle which individuates a subject's thoughts with respect to their cognitive profile – is in direct conflict with Publicity – hereby called 'the thought-transfer model of

[9] Löhr (2018) explicitly argues against conflating the philosopher's notion of concept with the psychologists'. Machery (2009) also makes clear that the psychologist's notion of concept, roughly bodies of knowledge that are retrieved by default in certain higher order cognitive tasks, is not at all the same as the philosopher's, so that these two fields are actually theorizing about distinct things.

communication'. I consider several ways out of the puzzle before conceding that the thought-transfer model is the weakest link.

Chapter 3 centers around the decades-old fight about whether Semantic Internalism is compatible with Publicity. Semantic Internalism is a fancy name for a familiar idea, i.e. that meaning is determined by use. More precisely, it is the thesis that the concepts we express are determined by our dispositional use of linguistic expressions. It is not hard to see that this view clashes quite directly with Publicity. Many seem to communicate well, and understand each other, while using representations in sharply distinct ways. I assess several ways to dissolve the conflict and end up suggesting a solution which I find conservative enough: weakening Publicity. Instead of proclaiming that communication (and understanding, and genuine agreement) require sameness of concepts between individuals, I argue that it only requires that the concepts individuals have are conceptually guaranteed to have the same extension. In other words, as long as the concepts we have are such that we could know that they co-refer just on the basis of our dispositional use of them, then they're good enough for communication, understanding, and genuine agreement.

Chapter 4 is about criteria for successful communication *per se* (as opposed to examining these criteria with the ulterior motive of individuating thought and concepts with them). The thread that runs through this paper is a series of variations on Loar-cases, thought experiments where subjects communicate, successfully or unsuccessfully, regardless of latching onto the same referential contents as each other. A key difference between some of the variations I consider is whether the subjects involved are ignorant or not about some identity fact involving the subject matter under discussion. I resist recent arguments by Unnsteinsson (2018) to the effect that successful communication requires that individuals lack false beliefs of a certain type. By doing this, I not only vindicate the possibility of understanding concomitantly with confusion, but I also conclude that communicative success cannot ever be reduced to a mere match of truth-conditions.

Chapter 5 is, again, a chapter based on a conflict. I consider a new trend in the philosophy of language and mind, the so-called 'relationist' views of thought and concepts, and argue

that it fails to guarantee a vindication of a substantial notion of concept publicity. It is, in summary, about the conflict between Relationism and Publicity. Relationism advokes individuating thought and concepts by means of external relations which token concepts may stand in. I assess several versions of these views, and try to argue that all of them commit the same two-fold sin: they fail to leave room for concept sharing between people who are not connected to each other, and they wrongly predict that some connected thinkers share their concepts even when their cognitive states are crucially different in more than one respect.

REFERENCES

Austin, John (1979) *Philosophical Papers, 3rd edition*, J. O. Urmson and G. J. Warnock (eds). Oxford University Press. First edition published in 1961.

Cappelen, Herman & Dever, Josh (2013) The Inessential Indexical. Oxford: OUP.

Frege, Gottlob (1892/1948), 'Sense and reference'. *Philosophical Review* 57 (3):209-230.

Frege, Gottlob (1918/1956), 'The Thought: A Logical Inquiry'. *Mind* 65 (259): 289-311.

Lewis, David. (1979) 'Attitudes De Dicto and De Se,' *Philosophical Review* 88: 513–543.

Loar, Brian (1976), 'The Semantics of Singular Terms'. *Philosophical Studies* 30 (6): 353-377.

Löhr, Guido. (2018) 'Concepts and categorization: do philosophers and psychologists theorize about different things?' *Synthese*. 10.1007/s11229-018-1798-4

Machery, Édouard. (2009). *Doing without concepts*. Oxford University Press.

Majid, A. (2010) Words for Body Parts. In: *Words and The Mind. How Words Capture Human Experience*. Ed. by Barbara C. Malt and Philip Wolff. Oxford University Press. Chap. 3, pp. 58 – 71.

Marques, Teresa & Wikforss, Asa Maria (eds.) (2020). *Shifting Concepts: The Philosophy and Psychology of Conceptual Variability*. Oxford: Oxford University Press.

Pagin, P. (2020) 'When does communication succeed? The case of general terms' in Teresa Marques and Åsa Wikforss (eds.), *Shifting Concepts*, OUP, (expected 2019).

Recanati, François (2012), *Mental Files*. Oxford University Press.

Schiffer, Stephen (1978), 'The Basis of Reference'. *Erkenntnis* 13: 171-206.

Unnsteinsson, Elmar (2018), 'Referential Intentions: A Response to Buchanan and Peet'. *Australasian Journal of Philosophy* 96 (3): 610-615.

THE PAPERS

1. What is Special about Indexical Attitudes?

Article published in *Inquiry*, 61:7, pp. 692-712, 2018.

Abstract

In this paper I examine the issue about whether indexical attitudes have any special properties or present any special challenge to a theory of attitudes and propositions. I will go over the claim that indexical cases coming from Perry (1977, 79) present no special challenges over and above more familiar cases of Frege's Puzzle before defending the claim that indexical attitudes are special in the sense that it is particularly hard to make sense of what is it for two people to believe the same thing *via* indexical expressions. In the end, I will assess Cappelen & Dever's (2013) considerations on intentional action and extract an argument from them that could, if successful, block the special indexical challenge. However, as I will go on to claim, their argument suffers from a considerable limitation and, in the end, gives us no overwhelming reason to believe that indexical attitudes are just as ordinary as any other.

1 Introduction: What exactly is essential about indexicality?

Sometime during the 1970's, philosophers started paying attention to how a set of attitudes, those which we normally express by means of indexical expressions, seem to pose a *special* challenge to theories of propositions. Two terminological remarks:

1. By "indexical attitudes" I mean those attitudes (such as beliefs and desires) that we normally express by means of indexical expressions (such as the first-personal pronoun). This is a neutral characterization and is silent about which special properties – if any – those attitudes are supposed to have.
2. By "propositions" I mean the semantically efficacious contents of attitudes by virtue of which those attitudes gain their normative and explanatory power (e.g. allowing us to explain subjects' actions, assess their rationality etc.).

The question that will concern me in this paper can be summarized as: do indexical attitudes present any *special* challenge to theories of propositions? The gloss on "special" is particularly relevant, since it might be that indexical attitudes are hard to characterize but only for the same reasons that other singular attitudes (e.g. those expressible by proper names) also are. Theories of proposition usually assume that

attitudes are *dyadic* relations between agents and propositions, which are then defined as *absolute* and *shareable* contents such that it is irrational to hold antagonistic attitudes (e.g. belief and disbelief) towards them at the same time. To be shareable is to be accessible by any speaker and to be absolute is to have a truth-value that depends only on the objective state of the world. There might be many arguments to the effect that those assumptions (and others) are not jointly consistent and Lewis might be right in saying that "the conception we associate with the word 'proposition' may be something of a jumble of conflicting *desiderata*" (Lewis, 1986, p. 54). Nonetheless, my only concern is finding out whether any of those arguments arise exclusively because of indexical attitudes.

In the following sections, we will see that demarcating the 'special indexical challenge' is far from an easy task. Indeed, many have failed to see that in order to prove that one such challenge really exists, it is not enough to show that one cannot fully characterize indexical attitudes neither by means of *de re* nor *de dicto* propositions, since the same predicament is true of other singular attitudes (section 2.1). However, I do think that at least one such challenge can be demarcated. It has got to do with the following question: what does it take for someone to share someone else's indexical attitudes? I will argue that the three most plausible answers to that question all lead to the rejection of some independently plausible thesis about propositions. Furthermore, I will show that this challenge is particular to indexical attitudes.

2 Frege's Puzzle and Indexical Cases: The "no *de re* and no *de dicto*" challenge

When one is trying to argue for there being something special about indexical attitudes, it is natural to turn one's attention to the work of John Perry (1977, 1979). Perry's memorable characters and thought-experiments quickly became part of the philosophical canon, although it is often not easy to tell what is their point supposed to be. I take it that Perry's discussion are usually based on two types of scenario. One type usually involves a subject who has information about himself without realizing that it is about himself. The key point of those cases is to show that indexical attitudes cannot be reduced neither to singular (*de re*) nor descriptive (*de dicto*) attitudes. In this section, I will argue that these "ignorance cases" do not display anything special about indexical attitudes. While

this is not an original point, I think that previous writers who have defended it failed to address certain worries. We will get to the second – and, in my opinion, more interesting – type of scenario in the following section.

The case of the amnesiac Rudolph Lingens is a paradigmatic example of an ignorance case about indexical attitudes (another famous example is the Messy Shopper):

> **Rudolph Lingens:**
> The amnesiac Rudolph Lingens is lost in the Stanford library. Lingens's amnesia is severe, and he has forgotten who he is. After reading a biography of Rudolph Lingens, he has a belief he could express by saying "Rudolph Lingens has been to San Sebastián." But at the same time, because he does not remember ever going to Spain, he does not have a belief that he could express by saying, "I have been to San Sebastián."[10]

The purpose of that kind of story is to show that, even for those who are exceedingly savvy about wordly matters, there is always room for some residual indexical ignorance (i.e. ignorance of matters that would have most naturally been expressed by means of indexicals). More particularly, Lingen's story suggests that, whichever proposition he expresses by[11] "Rudolph Lingens has been to San Sebastián", it is not irrational for him to endorse it while refraining to endorse whichever proposition he would have expressed by "I have been to San Sebastián". Thus, given the constitutive assumption that propositions should be individuated so that, if it is rational to hold antagonistic attitudes towards two of them, then they are not the same, it follows that: coming out from Lingen's mouth, those two utterances express distinct propositions.

At first, it is not obvious why that would be a particularly difficult challenge for a theory of propositions: we just need to search through the set of all propositions and assign these utterances two distinct ones. Since propositions are supposed to be the kind of thing that can be assessed for truth or falsity given a possible state of the world, they must either 'say' something general or particular about the world. To that fact

10 This is inspired by, but not identical to, Perry 1979, 21-2.
11 Wherever it does not lead to ambiguity, I will omit "uttering" when mentioning sentences with quotation marks. Thus, "The belief he would express by 'I am Rudolph Lingens'" is an abbreviation for "The belief he would express by uttering 'I am Rudolph Lingens'".

corresponds the distinction between descriptive (*de dicto*) and singular (*de re*) propositions. On the assumption that propositions are absolute, these two types of proposition are all the propositions we have.

But notice that no absolute proposition will do the job. Firstly, it cannot be the singular (*de re*) proposition that is true in every world in which Lingens has gone to San Sebastián, since this is most naturally seen as being the one expressed by the non-indexical utterance (and there has to be at least two distinct propositions around). It will also not do to characterize it as the proposition that is true in every world where the utterer of U [Lingens' utterance of "I have been to San Sebastián"] has been to San Sebastián, since it is easily conceivable that Lingens fails to realize the he himself is the utterer of U (and, as a consequence, believes that proposition without having the indexical attitude). Secondly, it seems that no purely descriptive (*de dicto*) proposition will be of any help. We could conceive of Lingens believing that, e.g. *the one and only amnesiac in the Stanford Library has been to San Sebastián* without him having any correspondent indexical attitude (because he could fail to believe the he himself is the only amnesiac around). This points generalizes: assuming that Lingens believe that he is in a reduplication world (where every qualitative property is satisfied by at least two different individuals), there will be no uniquely satisfied property F such that we could say that, when Lingens thinks of himself by means of the first-personal pronoun, he thinks of himself as the F. In other words, there is no property F such that it is irrational for Lingens to believe that he himself has been to San Sebastián while failing to believe that *the F* has been to San Sebastián.

This is our predicament: the logical space of absolute propositions is exhausted by the set of all *de re* and *de dicto* propositions and none of those serve to properly characterize indexical attitudes. Thus, these attitudes cannot be characterized as dyadic relations between agents and absolute propositions. However – and most relevantly for our concerns – is this predicament essentially related to indexical attitudes? It is quite easy to show that it is not.

It did not take many years until people realized that the "no *de re* and no *de dicto*" challenge, although legitimate, is less about indexicality than about the hyper-

intensionality of singular thought[12]. Both Magidor (2015) and Cappelen & Dever (2013) make a very strong case that this very challenge arises by means of cases which are *prima facie* unrelated to indexicals. To see that, notice how easy it is to construct an analogue of Linguen's story not involving indexicals (nor amnesia):

> **(non-indexical) Rudolph Lingens:**
> Rudolph Lingens is lost in the Stanford library. Lingens's knows that he is Rudolph Lingens but does not know that he is also known under a different name: Joseph K. After reading a biography of Joseph K (who happens to be himself), he has a belief he could express by saying, "Joseph K has a published biography." But at the same time, he also has a belief that he could express by saying, "Rudolph Lingens does not have a published biography."

This non-indexical story seems to pose the same challenge to theories of propositions than does its indexical counterpart. The difficulty in characterizing the propositions expressed by Lingens' two utterances is the same as we previously had. Firstly, one cannot characterize them by means of the *de re* proposition true in each world where Lingens has a published biography, since none of the two utterances seem to have a stronger claim on that proposition than the other. Secondly, no *de dicto* proposition (e.g. *the one and only lost person in the Stanford Library has a published biography*) can do the trick, since, for any *de dicto* proposition, we can conceive of Lingens believing it while disbelieving whichever are expressed by each of his utterances (and vice-versa). Thus, the predicament we reach is the same: proper name attitudes cannot be characterized as dyadic relations to absolute propositions.

Here is a reaction someone could have at this point: "even though the structure of the predicament is the same both for indexical and non-indexical cases, the solutions available to each are distinct – this is enough to show that there is something special about indexical attitudes". This is a fair claim and should be taken seriously. Which types of solutions can we employ after discovering that propositional attitudes cannot be characterized as dyadic relations to absolute propositions? It seems there are at least those

12 Stalnaker (1981) was probably the first to press this point. Cappellen & Dever (2013) brought these issues back to the spotlight and got the discussion running again. At the same time Magidor (2013) was advancing very similar claims.

two possible solutions: one can either reject the claim that attitudes are merely dyadic relations (and then introduce a third factor in their account) or the claim that all propositions are absolute.

Perry (1979) preferred to reject the dyadic claim than to let go of the absoluteness of propositions. Thus, his solution is a type of "third-factor strategy", according to which one needs a third ingredient – on top of the subject and the absolute proposition – to fully characterize an attitude. Perry called that third ingredient the 'belief state', although talk of 'guises' or 'modes of presentation' might also ring a bell. The important feature of that strategy is that it opens the way for there being different ways of, e.g. believing the same absolute proposition. This allows one to claim that, while Lingen's utterances (both in the indexical and in the non-indexical stories) have the same absolute *de re* proposition as their content, they encode different ways of believing that content. Perry's theory is an instance of the general strategy of differentiating between the *content* of a belief from *the way* it is believed.

The third-factor strategy applies across the board, i.e. there is nothing about the introduction of a third ingredient in an account of attitudes that seems to be specially about indexicals or which would preclude its application to non-indexical attitudes. For just one concrete example, this third ingredient could be the representational vehicle by means of which one believes a proposition. Thus, one could claim that it is possible to believe the same singular proposition *via* the proper name "Lingens", the proper name "Joseph K" or the first-personal pronoun "I" - and that those three manners of believing the same proposition all amount, in the end, to type-distinct beliefs. Whether this is a good theory or not should not concern us here. The important point is that, as a candidate solution to the phenomenon of opacity, the third-factor strategy does not seem to have a restricted application either to indexical or non-indexical attitudes.

However, things are not so clear when we look at the second general type of solution to the "no *de re* and no *de dicto*" predicament: the strategy of introducing non-absolute (or relative) propositions. Lewis' (1979) is perhaps its best example. Its crucial move is to complicate the notion of proposition. While one can assess the truth-value of an absolute proposition given no more than the full specification of a possible world, more is needed to assess a relative one. Take, for example, the case of first-personal attitudes. The natural idea is that a first-personal attitude expresses a proposition whose

truth-value varies across different subjects. According to this idea, the proposition Lingens expresses by "I have been to San Sebastián" is true or false *relative to a pair of a world and a thinker,* i.e. it is true just in case that thinker has been to San Sebastián in that world. Thus, it is true in the pair consisting of Lingens' world w and Lingens – the pair <w, Lingens> – but false in the pair consisting of our actual world w^1 and (up to this date) me – the pair <w^1, MV>. By introducing this new class of propositions, one is able to discriminate between having a first-personal attitude and having a proper name attitude. One can say that the former has relative propositions as its content whereas the latter has old-fashioned absolute ones. The same strategy can be generalized to other indexical expressions, e.g. temporal indexical attitudes ("Now is the time!") express time-relative propositions and locative indexical attitudes ("Here is the place!") express place-relative propositions. However, it is not clear whether that strategy sheds any light on non-indexical attitudes. That is, what relativization could we put in place so that one would be able to distinguish between the propositions expressed *via,* e.g. coreferential proper names?

Lewis (1979, p. 135) himself was the first to point out that his theory was a bit too specific. Interestingly, he did not think that this was a big problem: "My hunch is that this problem [the general phenomenon of singular thought] cuts across the issues I want to discuss [indexical attitudes], so I shall ignore it". (Lewis 1979, p. 135). Unfortunately for him, Lewis' hunch does not have as much weight these days as it had in the late seventies, when the literature on indexicality was flourishing and virtually no one doubted that indexical attitudes were special in some sense. Indeed, as soon as one starts to wonder whether indexical attitudes pose any special challenge over and above other singular attitudes, any theory of propositions which is only able to account for the former will need to have really good excuses.

I think the right reaction to have about those issues is simply to point out that Lewis' strategy can in fact be extended so as to encompass all singular attitudes. One just needs to take into account the so-called "centered descriptivist" strategy[13], according to which all singular expressions refer in virtue of being associated with definite descriptions containing indexical elements. Glossing over important details, one could

13 The roots of that theory are present in Lewis (1979, 1983) himself and Searle (1983). The epistemic two-dimensionalism of Chalmers (2006) and Jackson (1998) is a more recent instance of that general strategy.

claim that a name like "Lingens" refers to *the person being called "Lingens" by the persons from whom I've acquired that name* and that "water" refers to *the clear and potable liquid filling the rivers and oceans in our environment.* The presence of indexicals in these descriptions make it obvious that the attitudes characterized by them will not have absolute truth-conditions.

Now, even if centered descriptivism were proven to be a successful account of singular attitudes (we are as far as we can be from a consensus on that), one could complain that, as a solution to Frege's Puzzle, it only generalizes to non-indexical attitudes by reducing them to indexical ones. Thus, if it turned out that centered descriptivism is the best account of our singular attitudes, we would not have proven that indexical attitudes do not present any special challenge to theories of proposition: we would only have proven that much more attitudes are indexical in the first place[14].

Someone suspicious about the importance of indexical attitudes could then complain: "we have two extant solutions for the indexical and non-indexical cases of Frege's Puzzle – the Perrian and the Lewisian. They seem equally able to account for all the data but the latter additionally requires us to make the revisionary claim that all singular attitudes are indexical, thus, the Perrian one is clearly in better shape".

This leads us to a point in the discussion where not much is left to be said unless we are willing to get our hands dirty and start assessing concrete examples of Perrian theories and see how well they work. For just an example of how quickly things get complicated, it seems that concrete examples of Perrian theories typically end up having to claim that indexical attitudes have special properties not shared by non-indexical ones. Thus, even if the overall structure of Perrian solutions does not seem to imply any substantial distinction between indexical and non-indexical attitudes, what we find while examining concrete Perrian implementations is that, for one reason or another, they end up having to ascribe some special property to indexical ones. For just two examples, both Perry's own positive account and, more recently, García-Carpintero's (2016, p. 194) start from reasonable assumptions about attitudes and – because of issues related to action motivation – end up concluding that indexical attitudes have some form of 'limited accessibility', such that it is particularly hard to hold someone else's indexical attitudes.

14 Ninan (2016, p. 98) expresses the same worry.

Instead of assessing concrete accounts of singular attitudes and investigating whether there could be an implementation of a Perrian theory which did not entailed anything special about indexical attitudes, I will take the hint that it is in relation to action motivation and sharing-conditions that indexical attitudes really become peculiar and analyze the interconnections of those issues. My excuse for doing so is not so much that the first line of inquiry is impossible to be pursued, but just that, as long as there are interesting issues to dissect without leaving the most general level of discussion, it is important that it be done before going deeper into more intricate material. As I hope to show in the next section, one can advance a pretty robust argument to the effect that indexical attitudes are specially challenging without having to say anything substantial about which solutions to Frege's Puzzle one should adopt.

3 Indexicals and Action Explanation

As we've seen in the previous sections, the "no *de re* nor *de dicto* challenge" did not allow us to draw any fundamental distinctions between indexical and non-indexical attitudes. As far as it goes, we can only conclude that the attitudes we express by means of indexical expressions are usually not identical to the ones we express by means of proper names. One place to look for peculiarity of indexical attitudes is in its relation to intentional action. Minor tweaks to Lingen's first story may suggest that, not only his two utterances seem to express distinct propositions (or, alternatively, to express the same proposition *in a different way*) but also that only the indexical one is able to give rise to intentional action:

> **Rudolph Lingens (Action):**
>
> The amnesiac Rudolph Lingens is lost in the Stanford library. Lingens's amnesia is severe, and he has forgotten who he is. After reading a biography of Rudolph Lingens, he has a belief he could express by saying, "Rudolph Lingens's family is in San Sebastián" and a desire he could express by "I desire that Rudolph Lingens reunites with his family". However, only after realizing that he himself is Rudolph Lingens (a realization he could express by saying "I am Rudolph Lingens!"), does he acquire motivation to go out and book a flight to San Sebastián.

The underlying idea suggested by that version of the story is that, so long as one only has non-indexical attitudes, one will not be capable of finding out how the subject matter of these attitudes are related to oneself and, since performing an action requires knowing how one is related to the object of one's action, one will not be able to form an intention to act on the basis of them. In other words, one is never motivated to act unless one has some beliefs one would express indexically. Following Cappelen & Dever (p. 37) – henceforth C&D, one can rephrase that point in terms of what it takes to explain someone's intentional actions. The idea being that one cannot explain the intentionality of one's actions without mentioning, at some point, some indexical belief of that agent:

NIC (Non-indexical Incompleteness Claim):

All non-indexical action explanations/rationalizations are incomplete because of a missing indexical component.

A first point to note is how implausible NIC is given our ordinary practice of explaining the actions of our peers. Is it not the case that we, more often than not, explain the reason behind people's actions without mentioning any self-representational component? Is it not enough to explain why someone voted for the communist candidate to point out that this agent believed that if everyone voted for the communist candidate, then the world would be a better place? Why would we need to include any self-representational attitude in that action explanation? The belief she would have expressed by "I am a part of everyone", besides seeming silly, appears to play no role in her action. Perhaps the idea is that, in order to perform coarse-grained actions like voting, one needs to perform a multitude of finer-grained actions such as using one's hand to put the ballot on the urn. Then, one could claim that those basic bodily actions presuppose some sort of self-representation by means of the subject (e.g. knowing where one hand is in relation to the urn). But that claim is just as implausible as the idea we had begun with. As we perform basic bodily actions, there is very little need of representation to be going on in our conscience – we rarely have explicit thoughts about where our body is and its relation to the objects of our environment. As long as there is any use for the information encoded in those thoughts, it is something our subconscious motor system is more than capable of taking account of. In summary, there seems to be no argument to the effect that most

of our ordinary action explanations are incomplete or that every basic action need be motivated by some self-representation.

A second point to note is that, even if we grant that Lingen's action can only be explained by mentioning an "I"-belief ("I am Rudolph Lingens!"), NIC is a universal generalization and, strictly speaking, does not follow from considering an isolated case (C&D, p. 41-2). At most, Lingen's case would seem to prove a weaker claim such as:

NIC2 (Weak Non-Indexical Incompleteness Claim):

S*ome* action explanations/rationalizations ineliminably contain an indexical component.

However, C&D (p. 39) argue that there is no way "to read NIC2 as anything but a trivial corollary of the opacity of action explanations/rationalizations." Magidor (2013, p. 17) makes essentially the same point and observes how some non-indexical elements can *also* occur ineliminably in some action explanations. For example, in order to explain why Lingens acquires an intention to go to San Sebastián, one must mention certain beliefs and desires involving "San Sebastián", as opposed to "Donostía", even though both are names for the same city. Thus, it seems that, at least as far as this particular action explanation is concerned, the name "San Sebastián" occurs ineliminably.

I think C&D and Magidor's conclusions are a bit too quick. Even if NIC2 does not allow us to claim that every action is indexical (which I wholeheartedly agree is a hopeless claim to make), if we just ask ourselves about what does it take to share someone else's indexical attitudes and keep an eye on the implications of that question to action explanation, we will have in our hands a neat argument in favor of a special indexical challenge. Furthermore, I will have shown that this case follows from a premises that are accepted even by its most notorious detractors.

4 What does it take to share someone else's indexical attitudes?

So far we have conceded that indexical attitudes are sometimes not identical to proper name attitudes and even that indexical expressions might be ineliminable from some action explanations. However, none of this was enough to prove that indexical attitudes are *sui generis* in any substantial sense. I think there really is a special indexical challenge

in that vicinity, but the best way to get to it is *via* indirectly considering what does it take for someone to share someone else's indexical attitudes.

One important comment: in the remainder of that paper, I propose to focus on first-personal attitudes – those expressible by means of the first-personal pronoun. My reasons for doing so are multiple: most of the literature on indexical attitudes (e.g. C&D) focus only on issues arising from first-personal attitudes (indeed, the concept of *de se* attitudes is usually taken as synonym for indexical attitudes); even if temporally indexical attitudes could be proven to be as fundamental as first-personal attitudes, discussions about the former are additionally complicated because of issues in the metaphysics of time (e.g. it would seem that even deciding what "now" refers to depends on whether one is an A-theorist or B-theorist about time). Finally, even if one complains that my discussion is exclusively concerned with first-personal attitudes, if my argument is cogent, it is more than enough to show that indexical attitudes (or at least a subset of them) do indeed raise a special challenge to a theory of propositions. Without futher ado, the argument.

Take some arbitrary agent Amelia who has a belief she expresses by uttering "I am an aviation pioneer". What would it take for Berthold, who is distinct from Amelia, to hold the same belief that she expresses by means of that utterance? My argument, which I take to be a development of an argument found in Ninan (2016)[15], is that the three most plausible answers to that question each lead to the rejection of a different but independently plausible thesis about attitudes. Here are the three possible replies:

Option #1: For Berthold to share Amelia's first-personal attitude, he would need to form a belief (perhaps in response to Amelia's utterance) which he would express by means of "You [pointing at Amelia] are an aviation pioneer".

Option #2: For Berthold to share Amelia's first-personal attitude, he would need to have himself a first-personal attitude about himself, one which he would express by means of "I am an aviation pioneer".

Option #3: It is impossible to share someone else's first-personal attitudes (e.g. they are private, or limited accessibility etc.)

15 I think Ninan's paper has all the ingredients for the construction of this argument, but that it somehow fails to put all the pieces together. All in all, I was deeply influenced by reading his paper and see myself as developing its themes.

Let us begin by considering Option #1. It is supposed to be the most plausible of them, since it makes the sharing of a first-personal attitude a completely ordinary and easy affair. It can be seen as grounded on the principle that, when two subjects are in agreement with each other in virtue of some of their beliefs, then these beliefs are identical. In other words, since Berthold's second-personal belief seems to be the right belief to form in face of Amelia's utterance, it would seem that they agree with each other in virtue of the attitudes they express (even though she expresses it first-personally and he, second-personally). One who thinks that first-personal attitudes are just as ordinary as any others should be drawn towards Option #1, since it entails that thinking of oneself *via* the first-person is no more special than thinking about someone else *via* the second-person (and there does not seem to be anything mysterious about the latter). However, Option #1 has the consequence that the exact same belief will have distinct motivational roles for distinct agents. Imagine that Amelia and Berthold find themselves in a situation where there is urgent need for the expertise of an aviation pioneer, and that only Amelia fits that bill. Even though both believe the same thing by means of their, respectively, first-personal and second-personal attitudes, they would plausibly be disposed to perform different actions: Amelia would run to offer her help while Berthold would just stand by and hope for the best.

More generally, Option #1 would conflict with the principle (let us call it, following Ninan, "Explanation") that two agents who are doxastically identical should be disposed to perform the same actions. Explanation should not seem like a gratuitous *ad hoc* principle. Instead, it is one of the most entrenched principles governing folk-psychology. It is because of Explanation that it makes sense to explain people's behavior by means of their beliefs/desires and expect that this explanation be generalizable to distinct agents. If beliefs systematically had different motivational roles for different people, it would be impossible to predict people's actions based on what they believe and desire. It's not an exaggeration to say that, if Explanation were more often false than true, folk-psychology itself with its practice of ascribing semantic contents to attitudinal states would lose much of its *raison d'être*. Option #1 seems to entail that, for at least some attitudes, the first-personal ones, Explanation is bound to fail.

Since we have good reasons to protect Explanation, it could be good to try out the other possible answers to the attitudinal sharing question. However, both Option #2 and #3 seem to lead us to distinct conflicts with other independently plausible theses about attitudes and their contents. To be sure, both Option #2 and #3 allow us to maintain Explanation in its full generality, but that victory might be illusory seeing that they make us reject, respectively, the absoluteness of attitudinal contents or their shareability.

According to Option #2, when Berthold believes of himself that he is an aviation pioneer, he believes the same that is believed by Amelia when she believes of herself that she is an aviation pioneer. While it is true that two subjects who self-ascribe the same property are usually disposed to perform the same actions (unless they have other differing beliefs and/or desires in the vicinity), Option #2 entails that the same belief could be true for an agent and false for other (let us call "Absoluteness" the thesis it rejects). Thus, it entails that at least some objects of our attitudes are not absolutely truth-evaluable. Making way for objects of knowledge that are not themselves true or false irrespective of their knowers would be a significant departure from orthodoxy. Whether one could come up with good arguments for that is not what I intend to assess.

Option #3 obviously leads us to the claim that some attitudes are special in the sense that they cannot be shared by different subjects. Interestingly enough, Frege (1991/1918, p. 359) seemed to be attracted to such an account of first-personal thoughts, even going as far as claiming that "everyone is presented to himself in a special and primitive way, in which he is presented to no one else. And only [the thinker of a first-personal thought] can grasp thoughts specified in this way". I hope that my way of framing the special indexical challenge makes it clear why such a position would be particularly attractive: it allows one to characterize first-personal attitudes while maintaning Explanation and Absoluteness. However, its drawbacks are obvious. As soon as one rejects the general shareability of attitudes (let us call "Shareability" the thesis being rejected), one will need to come up with many revisionary stories about how communication works (since the same first-personal attitude cannot be put forward from speaker to hearer), disagreement (what is it to disagree about someone who claims to be an aviation pioneer?) etc.

In summary, whichever particular account of what sharing a first-personal attitude one chooses, one will need to reject some deeply entrenched thesis about

attitudes: Explanation, Absoluteness or Shareability. If, for example, one feels strongly about Explanation (as one should), then one will be led into either rejecting Absoluteness or Shareability. As Ninan (2016, 109-117) points out, that way of framing the dialectics fits really well with the fact that people like Frege and Perry were forced, for one reason or another, to claim that indexical attitudes are unshareable in some sense. The same goes for the fact that people like Lewis felt so strongly about introducing non-absolute propositions to account for our attitudes. These philosophers were choosing among the available routes given the special indexical challenge. Another plausible theoretical way out would be accepting that Explanation really fails for indexical attitudes. That is, one could claim that indexical attitudes are special in the sense that sharing them (in the sense of Option #1) does not entail being motivated to act alike. One could alleviate the consequences of that claim by suggesting that not identity of belief, but identity of type of indexical belief (e.g. when two subjects have first-personal beliefs about themselves) is the important relation for prediction of agency. If one chooses to take that route, the conclusion of the argument is that indexical attitudes are special in the sense that we need two distinct relations (as opposed to only one for the case of other singular attitudes) to characterize what is it to agree in virtue of them and to be disposed to act alike.

Finally, it should be clear that this is really a *special indexical* challenge. In other words, these complications would simply never arise if it weren't for indexical attitudes. That should be clear from the fact that, e.g. proper name attitudes, are such that it is very easy to characterize what it is to share them. We simply do not count someone who has a Lingen-belief as agreeing with someone who only has a Josef K-belief. More generally, we never count two people as agreeing about a belief when they use different proper names to express them. This ensures that there are no cases of agreement about proper name attitudes which motivate different actions. Thus, as far as proper name attitudes are concerned, we can just say that two subjects share a proper name attitude when they agree with each other in virtue of their respective attitudes.

5 Cappelen & Dever's Action Inventory Model

As convincing as the argument just presented is, it is not immune to criticisms. One way to put pressure on its conclusion is by challenging the claim that Option #1 is incompatible with Explanation. C&D's discussion of intentional action (p. 49-56)

suggests one ingenious way of arguing for that compatibility claim. In order to assess their argument, we need to first be clear on what is the precise formulation of Explanation in question. A first try would be:

(Explanation 1) If two agents have the same beliefs and desires, then, they will behave in the same way.

That goes in that right direction but not far enough. Firstly, notice that the antecedent of Explanation 1 is tremendously strong. Surely we do not need two agents to be doxastically identical in order for us to be able to predict that they will behave in the same way – what matters is that they have the same beliefs and desires about the relevant subject matter. For example, it does not matter that Amelia believes (and Berthold denies) that the Earth is flat, if the relevant subject matter is that people need an aviation pioneer. Secondly, it is also not the case that everytime two agents have the same relevant beliefs and desires, they will be able to behave in the same way. One can have as many attitudes as one likes, but if one's legs are tied down to the ground or if one is paralyzed by an evil genius, there will be many actions one will not be able to perform. Thus, it is natural to add some kind of *proviso* in the antecedent of the principle in order to account for cases where one is not able to behave as one desires because of external factors:

(Explanation 2) If two agents have the same *(relevant)* beliefs and desires *and the same (relevant) actions are available to them*, then, they will behave in the same way.

While that principle seems to fare better than the last one, it still lends itself to more than one interpretation. Notice that there is considerable vagueness about when two people behave in the same way or not. The question is: when are two action tokens instances of behaving in the same way? There are literally infinite way of classifying action tokens into action types and some of those ways will trivially entail that any two tokens were instances of the same behavior. For example, if one classifies Amelia's action as *an action performed by a human being*, then it was an instance of behaving in the same way as Berthold (even though she ran to help people and he just stood by wishing for the best). Conversely, if we individuate Amelia's action as *an action*

performed by Amelia, then it will never be the case that someone distinct from her will perform an action token that is an instance of behaving in the same way as her. Individuating actions is far from a trivial task, but I think we actually get by really well in our ordinary talk about actions and behavior. I think it is plausible to say that, given the description of ordinary scenarios such as Amelia and Berthold's, most people would be in agreement as to how to describe the actions they performed. Some discrepancy is surely to be expected, but I take it that most would lean towards saying that Amelia's action was something in the vicinity of *running to help people* and that Berthold's was that of *standing by*. Most importantly, these are descriptions of their actions which abstract away from their respective agents, focusing only in their qualitative component. In that spirit, let us assume that two agents can be said to have behaved in the same way if and only if they have performed action tokens which can be subsumed under the same ordinary (agent-neutral) action type. For example, Berthold would have behaved in the same way as Amelia if and only if he had performed an action which could be subsumed under the type *running to help people*.

That stipulation solves some of the interpretation problems, but notice that the *proviso* of Explanation 2 also talks about actions, so one should expect that the same complications about action individuation will also come about there. The question is: when can two agents be said to have the same (relevant) available actions? I take it that, ordinarily, we have no trouble assessing whether some action could have been performed by someone. It is easy to judge that someone in handcuffs could not perform the action of hugging, or that someone paralyzed by an evil genius could not jump around. I take this to show that in ordinary action-talk we have the following operative principle in the background: an agent can perform an action subsumable under an ordinary (agent-neutral) action type if and only if that agent can perform that action under that very same ordinary (agent-neutral) action type. For example, Berthold can perform an action subsumable under *running to help people* if and only if Berthold could have ran to help people.

(Explanation 3) If two agents have the same (relevant) beliefs and desires *and they could both have performed the (relevant) actions under an ordinary agent-neutral*

action type, then, they will perform action tokens which could be subsumed under the same ordinary agent-neutral action type.

So far so good. I am of the opinion that something like Explanation 3 comes very close to fully characterizing our ordinary folk-psychological principle equal behavior prediction for people who believe and desire alike. But notice that if this is the correct formulation of the Explanation principle, then it really is incompatible with Option #1. Both Amelia and Berthold, we have agreed, have the same relevant beliefs and desires (they both believe Amelia is an aviation pioneer and they both desire that an aviation pioneer run to help the people in need). They also seem to have the same relevant actions available, since Berthold could very well run to help people and Amelia could very well have stood by. There is no physical/psychological constraints that would incapacitate them from performing actions under those agent-neutral types. Nonetheless, they still go on to perform different actions: Explanation 3 fails. As we have seen in the previous sections, this is precisely one of the horns of the special indexical challenge: unless we are ready to admit that indexical attitudes have some special property (either the same indexical attitude has different motivational role for different agents, or they have non-absolute contents or they are unshareable), we reach a dead-end.

In order to rectify the principle, one can either accept that indexical attitudes are special and then tweak the principle so that it accounts for their particularities, or one can question Explanation 3 itself and try to reformulate its conceptual basis. I take it that the first option is implicitly adopted by folk-psychology: it is the reason why the thesis that indexical attitudes are special sounds so intuitive in the first place. C&D, not satisfied with that, are, to the best of my knowledge, one of the only authors to try to pursuit the second option[16]. By doing that, they end up making a fairly revisionary claim about intentional explanations, but if they succeed, that revisionary claim comes with the benefit of freeing us from the claim that indexical attitudes are more special than other ordinary ones.

Let us call the account which adopts Explanation 3 and that I take to be operative in ordinary action-talk "the indexical model". The point where C&D diverge from the indexical model is in their understanding of what does it take for an action to be available

16 Magidor (2015, p. 20) also goes in the same direction, but leaves her point in very broad strokes.

to an agent. As we have seen, there are infinitely many ways to individuate an action token. Notice that whether an action token is available to an agent highly depends on which specific action type one chooses to describe that token. If one describes the relevant action in Amelia and Berthold's case as *an action performed by Amelia*, then it trivially follows that Berthold is not able to perform it. However, that action type is a completely trivial one and surely not of the right granularity to be plugged into a principle of intentional explanation. Nonetheless, C&D believe that there are action descriptions which are, at the same time, (I) such that only Amelia could have performed it but not Berthold and (II) of the right granularity to be used in intentional explanations.

Take an *agent-specific description* of an action token to be a description which includes the ordinary (agent-neutral) component of that action but also its particular agent. In that sense, Amelia's action can be subsumed under the agent-specific description "that Amelia runs to help people". That type is not as trivial as the other one we've mentioned and could very well have the right amount of granularity to make our ordinary intentional explanations work:

(Explanation 4) If two agents have the same (relevant) beliefs and desires *and they could both have performed the (relevant) actions under their respective agent-specific action types,* then, they will perform action tokens which could be subsumed under the same ordinary agent-neutral action type.

As I read C&D, they are suggesting that Explanation 4 is just as good a principle for intentional explanations as Explanation 3. Furthermore, Explanation 4 seems to be perfectly compatible with Option #1. It is compatible because, for the case of Amelia and Berthold, its antecedent will come out false. That is, even though they share all the same relevant beliefs and desires, it is not the case that Berthold could have performed Amelia's action under the agent-specific type *that Amelia runs to help people*[17]. Naturally, the reason why he cannot perform an action token under that type is that he has no direct control over Amelia's body. Thus, according to Explanation 4, Amelia and Berthold would have both behaved in the same way if only they had the same actions

17 And it is also not the case that Amelia could have performed Berthold's actions under the type *that Berthold stands by wishing for the best*.

available, but they do not, thus, they perform distinct actions. This is the central claim of C&D's action inventory considerations and, in my opinion, one of their most important arguments against there being a special indexical challenge.

This, then, is the scenario we reach. Explanation 3 is the commonsensical formulation of the principle, but it fails for cases such as Amelia and Berthold's. Perry (1977, 1979) was precisely pointing to that fact when he discussed cases such as the Bear Scenario, which, just like our case, involves two agents who believe alike but act differently. The only way to hold onto Explanation 3 seems to be by claiming that, in these cases, certain indexical attitudes with special properties are playing some kind of special role. Explanation 4, on the other hand, is a bit revisionary about what it takes for an action to be available to someone, but it nonetheless seems to give the right result for this type of cases. If every case were like Amelia and Berthold's, then it would be a very easy victory for C&D and Explanation 4. But notice that for many other ordinary cases of intentional explanation, it is Explanation 4 that fails to output the correct predictions, while Explanation 3 works perfectly well:

(Nora's Case) Carlota knows that Nora is in danger and that if someone calls the police, then she will be saved. Since she desires that Nora be safe, she calls the police. Desmond also knows that Nora is in danger and that if someone calls the police, then she will be saved. He also desires that Nora be safe.[18]

I take Nora's Case to be a completely ordinary story about an action being motivated by some beliefs and desires. So much so that we do not even need to appeal to any indexical attitudes in order to explain why Carlota goes on to call the police. Now, it seems that any plausible principle of intentional explanation should predict that Desmond would also be disposed to behave in the same way as Carlota did. That is, any good account of our attitudes and the way they motivate action should be able to predict that, if Desmond also believes that Nora is in danger and that if someone calls the police, then she will be saved, then Desmond will in fact call the police. But notice that Explanation 4 is unable to give us that prediction. That can be quickly seen by observing that the

18 This is based on C&D's own case of Nora (p. 36-37), which has no significant difference from mine.

antecedent of Explanation 4 would only be satisfied if Desmond could have performed the action Carlota performed under the agent-specific type *that Carlota calls the police*. But surely Desmond cannot perform an action under that type for the same reasons given above: he does not have control over Carlota's agency. Thus, the antecedent of Explanation 4 is not satisfied for the case of Nora and that means that no prediction is made.

Contrast that with how well Explanation 3 fares for that case. Since Desmond could very well perform Carlota's action under the type *calling the police*, we get it that the antecedent of Explanation 3 is satisfied. Thus, it correctly outputs the prediction that Desmond, because he shares Carlota's actions and has the same relevant actions available, will also call the police.

So things were not so favorable to C&D's principle in the first place. While Explanation 4 seems to be compatible with Option #1, it does not allow us to make the right amount of predictions of behavior for our concepts of belief and desire to play the role that they have in folk-psychology. In other words, Explanation 4 allows us to go on without special indexical attitudes makes us unable to make very ordinary predictions of same behavior. Explanation 3, on the other hand, allows us to make those ordinary predictions, such as in the case of Nora, but it does get into trouble when considering cases such as Amelia and Berthold's. It is because of those cases that indexical attitudes are ascribed such special properties as unshareability of non-absoluteness. Now, what C&D would need to show – if Explanation 4 is to be considered a real contender for principle of intentional explanation – is that they can tweak their theory in such a way as to be able to explain simple cases such as Nora's. My own suspicion is that the only way available to them will be claiming that some attitudes have some kind of special property and these special properties explain why things are different from cases such as Amelia and Berthold's to Nora's one. But if they are willing to admit that some attitudes have special properties, then we might as well just stick to the indexical model with its special indexical attitudes. That model is not only well equipped to deal with intentional explanation but also in consonance with our folk-psychological intuitions about how beliefs and desires work.

6 Conclusion

In this paper I have examined the issue about whether indexical attitudes have any special properties or present any special challenge to a theory of attitudes and propositions. In section 2, I defended the thesis that, as far as some cases such as the Messy Shopper go, they do not present any special challenge over and above the general challenge of Frege's Puzzle, which is common to all singular attitudes, not only indexical ones. In section 3, I proposed to focus on the interrelations between indexical attitudes and action explanations. I argued that, while it is implausible to claim that all action must be motivated by indexical representations, that at least some of them might be, and that this could prove to be an interesting point in itself. In section 4, I built on previous considerations and showed how one can get to a special indexical challenge by asking what does it take to share someone else's first-personal attitudes. I defended the thesis that every plausible answer to that question leads us to reject independently plausible theses about attitudes. I argued that this is a robust defense of the claim that indexical attitudes have special properties and that it is based on very plausible premises. Finally, in section 5, I assessed Cappelen & Dever's considerations on intentional action and extracted from their work an argument which could block the special indexical challenge offered in the previous section. I argued against their argument on the basis of the fact that their account of intentional explanations fail to output the correct predictions for many ordinary cases. On the other hand, what I called the indexical model – which has in its core the claim that indexical attitudes are special – seems well-suited to account for intentional action of all varieties. Thus, I concluded that these authors have failed to provide us with a reason to stop believing that there is really something peculiar about indexical attitudes.

References

Burge, Tyler (2000). Reason and the first person. In C. Wright, B. Smith & C. Macdonald (eds.), *Knowing Our Own Minds*. Oxford University Press.

Cappelen, H. and J. Dever (2013) *The Inessential Indexical*. Oxford: OUP.

Chalmers, D. (2006) The Foundations of Two-Dimensional Semantics, in *Two-Dimensional Semantics: Foundations and Applications*, M. Garcia-Carpintero and J. Macia (eds.), Oxford: Oxford University Press, pp. 55–140.

García-Carpintero, Manuel & Torre, Stephan (eds.) (2016). *About Oneself: De Se Thought and Communication*. Oxford University Press.

García-Carpintero, Manuel. Token-Reflexive Presuppositions and the De Se. In Stephan Torre & Manuel Garcia-Carpintero (eds.), *About Oneself: De Se Thought and Communication*. Oxford University Press.

Frege, Gottlob (1991). Collected Papers on Mathematics, Logic, and Philosophy. Wiley-Blackwell.

Jackson, F. (1998), *From Metaphysics to Ethics: A Defence of Conceptual Analysis*, Oxford: Oxford University Press.

Lewis, D. (1979) 'Attitudes De Dicto and De Se,' *Philosophical Review* 88: 513–543.

Magidor, Ofra (2015). The Myth of the De Se. *Philosophical Perspectives* 29 (1):249-283.

Ninan, Dilip (2016). What is the Problem of De Se Attitudes? In Stephan Torre & Manuel García-Carpintero (eds.), *About Oneself: De Se Thought and Communication*. Oxford University Press.

Perry, J. (1977). Frege on demonstratives. *Philosophical Review* 86 (4):474-497.

Perry, J. (1979) 'The Problem of the Essential Indexical,' *Noûs* 13: 3–21.

Searle, John (1983). *Intentionality*. Oxford University Press.

Stalnaker, Robert C. (1981). Indexical belief. *Synthese* 49 (1):129-151.

2 A Puzzle about Understanding

(with Andrea Onofri[19])

Abstract

It seems plausible that understanding our peers in communication requires entertaining the same thoughts as they do. We argue that this view is incompatible with other, independently plausible principles of thought individuation. Our argument is based on a puzzle inspired by the Kripkean story of Peter and Paderewski: having developed several variations of the original story, we conclude – albeit only tentatively – that understanding and communication cannot be modeled as a process of thought transfer between speaker and hearer. While we are not the first to reach this conclusion, the significance of our argument lies in the fact that it only relies on widely accepted premises, without depending on any especially controversial theory of mental and linguistic content. We conclude with a plea: if understanding doesn't require thought identity, then we urgently need a systematic account of that relation which allegedly explains communicative success without amounting to full identity of thought.

1 Introduction

What conditions must be satisfied for a hearer to successfully understand a declarative utterance? This is obviously a central question for any account of linguistic communication. Unsurprisingly, the question has received a lot of attention in recent and not-so-recent discussion within philosophy of language, philosophy of mind, and other areas. Also unsurprisingly, no universal consensus has been reached, yet there are some basic principles which many philosophers working on these topics have accepted as *prima facie* plausible, maybe even indispensable – "constraints"[20] that any plausible account in this area should meet.

Our first goal is to identify a set of principles that have enjoyed this privileged status and show that they cannot all be true. We will present a puzzle about understanding: having examined a set of attractive and widely accepted claims about thought and communication, we will show that the set is inconsistent, with the aid of some plausible auxiliary premises. We also have a second, more constructive goal: offering a tentative solution to the puzzle. Our ultimate diagnosis is that an influential and popular account of communication – the so-called "thought-transfer model" – ought

[19] Andrea is a *profesor de tiempo completo* (tenure-track) in the Philosophy programme at the Universidad Autónoma de San Luis Potosí, Mexico.

[20] We borrow the expression from Fodor (1998).

to be rejected as the least plausible element of our inconsistent set of principles. While we're not the first to suggest that this commonsensical view of communication must go, our argument, as will be seen, is more general – and thus more definitive – than others previously proposed. At the same time, we will show what this theoretical choice involves and why it should not be taken lightly – which is why our solution is only a tentative one.

Other authors have presented arguments in the vicinity of our puzzle. The basic idea takes inspiration from Kripke (1979) and appears in Loar (1988) and Onofri (2016); to a lesser extent, Crimmins (1992) and Heck (2002) also present a similar argument. While we are obviously indebted to these predecessors, it is also important to highlight the differences. We will discuss a number of different cases where our puzzle arises. These cases are both different from and more varied than those discussed by the authors cited above. This is necessary, for some alleged solutions to the problem will turn out to be insufficiently general – they can deal with some, but not all the variations of the puzzle. So the argument only has force when we consider the full range of cases, something that the above authors do not do. Furthermore, all of the authors we have just cited present the problem as a conflict between the thought-transfer model and Fregean criteria of thought individuation. But this is incorrect, or at least incomplete, since other important premises are needed to generate the puzzle. We will offer a more complete view of the territory – first by making explicit all the principles that form our inconsistent set, then by discussing in detail which of those principles ought to be rejected.

The paper has the following structure. In Section 2 we introduce the basic case on which our argument is based. In Section 3 we explain why the case in question presents us with a puzzle. In Section 4 we devise two variants of the basic case to show that the puzzle is even more pressing than it initially seems. In Sections 5 and 6 we consider two attempts to solve the problem by dropping some influential principles about thought preservation, communication, and individuation. Having rejected these strategies, we propose our solution to the puzzle in Section 7, where we identify the thought-transfer model as the weaker element of our inconsistent set and explain why our diagnosis of the problem is only a tentative one.

2 Considerations on a well-known Kripkean theme

In a famous paper from 1979, Kripke tells the story of Peter's unfortunate predicament. As Kripke describes the story (Kripke 1979, pp. 398-399), Peter first learns of this man called 'Paderewski' who is an accomplished pianist, thus acquiring the belief that he would express as: 'Paderewski has musical talent.' Later, in a different context and surrounded by different people, Peter learns of somebody called 'Paderewski' who was a politician in pre-war Poland. Since Peter doubts that politicians ever have musical talent, he infers that this man is not the one he had previously learned about and that it is just a coincidence that they are namesakes. Peter thus acquires the belief that he would – in this context – express as: 'Paderewski does not have musical talent'. The problem, however, is that Peter is mistaken: contrary to what he thinks, the "first" and the "second" Paderewski are the same person. So Kripke's question arises: does Peter, or does he not, believe that Paderewski has talent? He surely has two beliefs with incompatible truth-conditions, one which he'd express by 'Paderewski has musical talent', another which he'd express by 'Paderewski does not have musical talent'. Should we then conclude that Peter believes a contradiction? Kripke never suggests a definitive answer, contenting himself with cryptic remarks such as "[when we consider these cases] our practices of interpretation and attribution of belief are subjected to the greatest possible strain, perhaps to the point of breakdown" (Kripke 1979, p. 269). Yet these cases can still yield fruitful philosophical reflection, even if we set aside Kripke's question about Peter's beliefs. More specifically, we suggest reframing the Kripkean case as raising a problem about successful understanding and communication, rather than belief attribution. As we will see, while reporting Peter's beliefs may be particularly tricky, the issue of whether subjects like Peter can engage in successful communication seems much more tractable.

Before presenting our variants of Kripke's case, it will be helpful to introduce some basic terminology that is often used in the literature.[21] Suppose speaker *S* makes an utterance directed at hearer *H*. We can then say that *S expresses* a certain *thought* t^S through her utterance, while *H entertains* a certain thought t^H as a result of the utterance. If *H* believes the utterance is true, then she stands in the *belief-relation* with thought t^H; if *H* believes the utterance is false, then she stands in the belief-relation with the negation

[21] See for instance Frege (1892, 1918/1956) and Fodor (1998).

of t^H. Using an equivalent formulation, we can say that *H* can *accept* t^H as true or *reject* t^H as false. Finally, *H* might be agnostic and suspend judgment about the utterance, in which case she will simply entertain t^H without accepting it nor rejecting it. We take this terminology to be minimal enough that theorists of different persuasions could accept it – for instance, we are not taking any particular stance on whether thoughts are abstract objects[22] or mental representations[23], nor are we subscribing to any particular theory of belief ascription[24]. We also take no stance on whether thoughts are wholly individuated by their semantic properties or by non-semantic features as well – such as syntactic properties in the Language of Thought, origin, causal-historical relations to external objects, and so on. Our goal is to describe our cases and set up our discussion in a way that is compatible with different theoretical frameworks.

Let's now introduce the basic version of our case. The central idea is simple: we imagine the confused subject Peter trying to communicate with a non-confused subject – that is, someone who is aware of the relevant identity. For instance, suppose Wendy knows that the pianist called 'Paderewski' and the politician called 'Paderewski' are the same person. Now imagine that Wendy and Peter meet in a context (C^1) where classical music is a salient topic, such as a concert where Paderewski has just given a performance. Wendy utters 'Paderewski has musical talent', thereby expressing a thought *t*. Given the context of Wendy's utterance, Peter thinks she must be talking about the pianist and takes her utterance to be true. So there is a certain thought t^1 that Peter accepts – a thought he would express by 'Paderewski has musical talent' in contexts where the pianist was under discussion. Think now of a distinct context (C^2) where pre-war politics in Eastern Europe is the main topic, such as a rally where Paderewski has just given a political speech. Wendy utters 'Paderewski has musical talent', thereby expressing the same thought that she expressed in her previous conversation with Peter. In the new context, however, Peter thinks that Wendy must be talking about the politician and takes her utterance to be false. So there is a certain thought t_2 that Peter rejects – a thought he would express by 'Paderewski has musical talent' in contexts where the politician was under discussion.

[22] See for instance Frege (1918/1956), Peacocke (1992).

[23] See for instance Fodor (1998), Laurence and Margolis (1999, 2007).

[24] The literature on this particular topic is very large – for some fairly recent works, see for instance Braun (1998), Soames (2002), Chalmers (2011).

What should we say about Peter and Wendy's communicative exchanges in C^1 and C^2? Should we say that their communicative efforts are successful? Put more simply, does Peter understand Wendy? An affirmative answer seems *prima facie* plausible. In C^1, Peter correctly identifies the referent as the pianist called 'Paderewski'; in C^2, he correctly identifies the referent as the politician called 'Paderewski'. Peter has thus identified the object of the utterance in both cases, and he has done so in a way that does not seem "deviant" or "lucky".[25] This suggests that Peter understands Wendy's two utterances of 'Paderewski has musical talent'. However, this seemingly innocent assessment of the case can quickly lead to troubling consequences.

3 The puzzle

A widely endorsed principle holds that, whatever else one might want to say about thought individuation, it must closely track the thinker's rationality, so that it is irrational to take contrasting attitudes towards a thought t^A and a thought t^B if t^A and t^B are the same thought. In converse form, the claim is: if a subject rationally holds contrasting attitudes towards a thought t^A and a thought t^B, then t^A and t^B must be distinct. In the literature this principle goes by various names; following Schiffer (1978), we'll call it 'Frege's Constraint' because of its strong connection to Frege's seminal ideas in 'On Sense and Reference' (Frege 1892).

Now, the case described in the previous section is precisely the type of situation where Frege's Constraint applies. As Kripke himself notes, it seems incorrect to describe Peter as irrational – Peter might be a leading logician and thoroughly desire to avoid contradiction, while still holding incompatible beliefs because of his ignorance about the relevant identity. This is where Frege's Constraint comes in: if Peter is rational, then the relevant thoughts must be different. To explain this in more detail, consider context C^1 first. Since Peter thinks the pianist is under discussion, he takes Wendy's utterance 'Paderewski has musical talent' to be true. So there is a certain thought t^1 that Peter accepts as true. Now consider context C^2: here, Peter thinks the politician is under discussion, so he takes Wendy's utterance 'Paderewski has musical talent' to be false. So there is a certain thought t^2 that Peter rejects as false. Peter's attitude towards t^1 is

[25] For an example of a case where the hearer identifies the right referent in a lucky way and therefore does not seem to understand, see Loar (1976).

therefore different from his attitude towards t^2: Peter accepts t^1 as true, whereas he rejects t^2 as false. By Frege's Constraint, it follows that t^1 and t^2 are different thoughts. An attractive and popular principle of thought individuation thus demands that we take Peter's thoughts as distinct.

Let's now introduce a second influential principle: the thought-transfer model of communication[26]. This model is based on a simple idea appearing in the work of philosophers as distinct (and distant) as Locke, Frege and Stalnaker[27]: if you make an utterance and I understand you, then I must be entertaining the same thought that you expressed. Understanding, in other words, requires that speaker and hearer entertain the same thought. Successful communication presumably requires understanding, so a speaker and a hearer who communicate successfully must entertain the same thought. The attractiveness of this simple thought-transfer model of understanding and communication is easy to see. How could I be said to have understood you if the thought I have entertained is distinct from the one you expressed? And how could you have successfully communicated your thoughts to me if your thoughts – the ones you expressed – are not the ones I have entertained?

We can now formulate the main conflict this work is about. Return to Wendy and Peter's communicative exchanges in C^1 and C^2. As we have seen, Peter seems to understand Wendy in both of those exchanges, yet this claim seems incompatible with the two influential principles we have just introduced – namely, Frege's Constraint and the thought-transfer model. To see why, note that, if Peter understands Wendy in C^1, then by the thought-transfer model he entertains a thought t^1 that is the same as Wendy's thought. *Mutatis mutandis*, if Peter understands Wendy in C^2, then by the thought-transfer model he entertains a thought t^2 that is the same as Wendy's thought. So, if Wendy expresses the same thought in those two conversations, that thought will be identical with each of Peter's thoughts t^1 and t^2. By the transitivity of identity, Peter's t^1 and t^2 will then be the same thought; but that is incompatible with Frege's Constraint, which holds that Peter's thoughts are distinct. Here is a more precise formulation of the argument – what we'll call 'the puzzling argument' or more simply 'the puzzle':

[26] See for instance Egan (2007) and Torre (2010), who call it the 'belief-transfer model of assertion'.
[27] For a recent defense, see Cumming (2013) and Onofri (2018).

The puzzle

1. Wendy expresses the same thought *t* by uttering 'Paderewski has musical talent' in C^1 and C^2.
2. Peter understands Wendy's utterances in C^1 and C^2.
3. As a result of Wendy's utterances in C^1 and C^2, Peter entertains a thought t^1 (in C^1) and a thought t^2 (in C^2) which are both identical with Wendy's thought *t* [by 1, 2, and the thought-transfer model].
4. t^1 and t^2 are identical [by 3 and the transitivity of identity].
5. t^1 and t^2 are distinct [by Frege's Constraint].

(4) and (5) are contradictory, so something must have gone wrong – but what? Several strategies are available. One could reject premise (1) by arguing that Wendy does not express the same thought in C^1 and C^2. One could reject (2), claiming that Peter does not understand Wendy in one or both contexts. One could reject the thought-transfer model and thereby block the step from (1)-(2) to (3). Finally, one could reject (5) by rejecting Frege's Constraint.[28] We'll discuss each of these possible solutions in what follows.

4 Does Peter understand? Two further cases

One could simply rule out our initial assessment of the case as incorrect, holding that people who are as confused as Peter do not get to successfully understand non-confused people like Wendy. Indeed, authors like Cumming (2013) and Unnsteinsson (2018) have defended views in the vicinity. The key move here would be emphasizing the gravity of Peter's confusion. One might begin by noting that Peter will not draw some of the inferences that Wendy expects him to draw: for instance, hearing Wendy's utterance in C^2, Peter will fail to infer that there's a pianist who is also a politician. One might then note that Peter's mistake is even more serious than just failing to draw some inferences, for he is in fact taking a single name to be two. In C^2, Peter fails to recognize that the name Wendy is using, 'Paderewski', is the same name that she had used in a previous context – Peter is failing to correctly individuate the words his interlocutor is employing.

[28] For now, we set aside premise (4) and the transitivity of identity, but we'll return to this in Section 7.

So Peter has failed to understand Wendy and premise (2) of our puzzling argument is false.

Now, we are unsure whether Peter misunderstands in this case – he has certainly made a mistake, but he has also identified the referent correctly. Furthermore, his mistake is not so uncommon. Many subjects "take one thing to be two", so if Peter misunderstands this might overgenerate: many communicative interactions which we would intuitively take to be successful would have to be classified as unsuccessful.

Luckily, we need not delve deeper into the issue. We will now discuss two variants of Peter's case: in these variants it seems clear that Peter does understand Wendy, so that it would be highly implausible to reject premise (2) of the puzzling argument. We will therefore need a different strategy if we want to solve the puzzle in its full generality.

4.1 First variant: indexicals

The puzzle from the previous section involved proper names, but the same problem arises with other kinds of expressions. Consider for instance the following variant, which involves the first-person pronoun. At a party, Peter is talking to a group of guests. The guests are talking about their respective talents and one man says: 'I have musical talent'. Peter trusts the guest and accepts the utterance as true. Just a few hours later, the same man is giving a speech at a political rally. Peter is also attending, but – being far from the stage – he does not recognize the speaker as the man he met at the party. While boasting about his own qualities, the speaker says: 'I have musical talent.' Peter is convinced that the speaker must be lying or exaggerating, so he rejects the utterance as false.

As in previous cases, Peter is confused – having failed to recognize the speaker as the party guest, he takes one thing to be two. The crucial question then is: has Peter understood the man's two utterances in spite of being confused? It seems clear that he has. Both at the party and at the rally, Peter identifies the referent correctly – at the party, he identifies the referent as the guest in front of him; at the rally, he identifies the referent as the man on stage. Furthermore, being a competent speaker, Peter knows very well that the same type of expression (the first-person pronoun) is being used in the two cases. This constitutes a difference with cases involving proper names, where Peter takes one name ('Paderewski') to be two. Of course, Peter has made a mistake: he failed to

recognize the man, so he now holds two incompatible beliefs about him. But this mistake cannot possibly impair his understanding of the utterance; if it did, then communication would fail whenever the first-person pronoun was being used by a speaker that we have failed to recognize.

This variant is of course modelled on the case examined in the previous section. One might suspect that the reason why we have been able to reframe the puzzle has something to do with the peculiarity of the first-person indexical and its mental analogue: the self-concept. Much has been written about the *sui generis* character of first-person thought, so one may suspect that any argument based on this puzzling type of representation cannot be so quickly applied to a wider range of cases. Indeed, one popular view, stemming from Frege (1918), is precisely that first-person thoughts are not shareable – they are of limited accessibility, only thinkable by their owners (Perry 1979). If one is sympathetic to this view, one will hardly be moved by the present argument.

But the argument just presented does not at all depend on the first-person indexical – it can very easily be reformulated with other context-sensitive expressions, like the second-person pronoun. Again, suppose Peter is at a party, talking to a group of guests. Wendy is also present and at one point she says to one of the guests, a man: 'You have musical talent.' Peter trusts Wendy's judgment and accepts her utterance as true. A few hours later, the same man is giving a political speech. Peter and Wendy are in the audience; Wendy knows that the speaker is the man from the party, but Peter fails to recognize him. At some point Wendy says to the speaker: 'You have musical talent.' This time, however, Peter thinks that Wendy must be wrong, so he rejects her utterance as false. This case does not involve the first-person indexical, but the same problem arises: through a perfectly ordinary interpretation process, Peter has assigned the right referent both to Wendy's first utterance (the man at the party) and to her second utterance (the man on stage). It thus seems clear that Peter understood both utterances. One could easily construct further variants of the puzzle with other indexicals, but this case is enough to prove our point: the puzzle does not depend on any peculiar features that the first-person indexical might have.

It will be helpful to explain in more detail how the argument unfolds in cases involving indexicals. Here is the structure of the puzzle in the last case we proposed – the one involving the second-person pronoun:

The puzzle (indexical variant)

1. Wendy expresses the same thought *t* by uttering 'You have musical talent' in C^1 and C^2.
2. Peter understands Wendy's utterances in C^1 and C^2.
3. As a result of Wendy's utterances in C^1 and C^2, Peter entertains a thought t^1 (in C^1) and a thought t^2 (in C^2) which are both identical with Wendy's thought *t* [by 1, 2, and the thought-transfer model].
4. t^1 and t^2 are identical [by 3 and the transitivity of identity].
5. t^1 and t^2 are distinct [by Frege's Constraint].

Our puzzle thus turns out to be more general – and more difficult – than one might have thought at first. It is not only a puzzle about names, for it also arises with indexicals; it is not only a puzzle about the first-person, for it also arises with other indexicals; and it cannot be solved just by claiming that Peter misunderstands, for that solution is completely implausible in the cases we have just examined.

4.2 Second variant: Wendy knows

It gets worse: there are other variants of the puzzle where there is no reason to claim that Peter misunderstands. Suppose Wendy is well aware of Peter's confusion: she knows that the pianist and the politician called 'Paderewski' are one and the same person, but she also knows that Peter is not aware of their identity. Then Wendy says 'Paderewski has musical talent' in C^1. She knows that Peter will take her utterance to refer to the pianist called 'Paderewski', but not to the politician; however, she thinks it is not important for her communicative purposes to correct Peter's mistake, so she does not. Later, in C^2, Wendy says: 'Paderewski has musical talent'. Again, Wendy knows that Peter will interpret her utterance as referring to the politician, not the pianist, yet she does not bother to correct Peter's mistake, thinking that this will not interfere with her communicative purposes. The question then is: has Peter understood Wendy's two utterances?

It seems clear that he has. Wendy is well aware that Peter will not take her utterance to refer to both the pianist and the politician, yet she does not bother to correct him, thinking that he'll understand her anyway. In this new setup, it's hard to see why we should refrain from ascribing understanding to Peter. Wendy thinks that Peter understood her despite his confusion; why should we think otherwise?

We thus have a further variant of the puzzle where it is very clear that Peter understands Wendy.[29] In combination with the indexical case presented above, this shows that there are at least two kinds of cases where it is wildly implausible to reject premise (2) in order to solve our puzzle. Let us then turn to examine other possible solutions.

5 Belief retention and communication

One strategy that might look promising at this point is rejecting premise (1). Why think that the speaker is expressing the same thought through her utterances in C^1 and C^2? The answer is that (1) is supported by some very plausible principles concerning the retention and communication of belief.

Start with belief retention. Ordinary subjects like Wendy seem able to retain their beliefs across time. In all of our cases, there is a belief that the speaker retains throughout her two exchanges with Peter. Consider Wendy, who asserts 'Paderewski has musical talent' in C^1 and C^2. At some point, Wendy formed a belief about Paderewski's talent, a belief she expressed in C^1 and then expressed again in C^2. In her second conversation with Peter, Wendy intends to re-express a belief that she holds from before, not a "new" belief she formed after their first conversation.

This natural and appealing description of the case assumes that the same belief was formed and then retained – there hasn't been a succession of *distinct* beliefs, but rather a *single* belief that persisted across time. This claim is not just pre-theoretically plausible; the idea of belief retention also allows us to distinguish two very different ways in which a subject might think about the same object at different times. One can think about the same object without representing it *as* the same,[30] failing to recognize it as the

[29] The argument in this version would have the same structure as previous ones, so we won't repeat it here.
[30] For discussion of the distinction see Fine (2007), Recanati (2012).

object that she was previously thinking about. John Perry provides an example of this phenomenon:

> Suppose I ask two friendly Chicagoans which building is Union Station. One points to a building the three of us are standing beside, the other to a building some distance away, across a street. Both are correct, as Union Station is mainly under the street, rising up on either side of it. But I do not know this. My mind changes as I turn to each of the honest-looking natives. First I accept 'This is Union Station,' and then 'That is Union Station.' (Perry 1980, p. 323)

As Perry notes, 'this does not seem to be a case of what we would ordinarily call continuing to believe the same thing.' (*ibid.*) The subject in Perry's example ascribes the same property to the same object at both times, but her two beliefs at those times are not connected in the right away, for he (wrongly) takes himself to be thinking about different objects. We can therefore describe the subject as having formed a new belief. Contrast this with someone like Wendy, who ascribes the same property to the same object at different times *without* thinking that the referent has changed: this subject is correctly described as having retained the same belief. In light of all this, premise (1) of our argument becomes very plausible. Consider the basic version of our case (analogous considerations apply to the two variants from Section 4). Wendy knows there is a pianist called 'Paderewski' who is also a politician. At some point, she forms a belief about Paderewski – she now believes that the pianist and politician called 'Paderewski' has musical talent. In her first conversation with Peter at the concert, she expresses that belief through her utterance: 'Paderewski has musical talent'. Later, in her second conversation with Peter at the political rally, she expresses that belief again by uttering the same sentence: 'Paderewski has musical talent'. In light of the above considerations about belief retention and communication, we take this to be a correct representation of Wendy's situation. So Wendy has expressed a single belief in C^1 and C^2, and premise (1) is true.

Could one still argue against this premise? One possible counter-argument has to do with the subject's knowledge of the relevant identity. To see how this might go, start with a basic question: if a subject thinks that *a* is different from *b*, but later learns that *a*

and *b* are the same object, what happens to her mental representations of *a* and *b*? Does the subject "merge" the two representations into one, or does she keep the two representations as separate items in her cognitive repertoire?[31] Each of these two options has been defended in the literature; here we will not discuss their respective merits, but only a possible connection with our main topic. Suppose we reject the "merge" model, holding instead that someone who learns the identity between *a* and *b* will keep two separate representations – for instance, she might keep two distinct "files" for the object, rather than merging them into a single file (see Recanati 2012, 2016). One could then reject premise (1) of the puzzling argument, arguing along the following lines: the speaker in our cases has two separate mental representations for the referent, using one or the other depending on the context. For instance, even though our speaker Wendy knows that the pianist and the politician are the same person, she might still have two separate files for Paderewski. When the conversation focuses on Paderewski's musical skills, Wendy would use the first file; however, Wendy would use the second file when having a conversation about Paderewski's political activity. According to this argument, Wendy does not express the same thought in her two conversations with Peter, but rather different thoughts constituted by different mental representations of Paderewski. So the first premise of our puzzling argument is false.

Unfortunately, this response assumes that Wendy was – at some point in the past – ignorant in the same way that Peter currently is, being unaware that the pianist and the politician are the same person. But suppose Wendy *always* knew about Paderewski's two talents; then there would be no moment in the past where she had two distinct files about him (one for the pianist, the other for the politician). The question about whether files are merged or not would then be out of place. This is even clearer in other versions of the puzzle. Suppose our speaker makes two utterances of 'I have musical talent' at different times (see Section 4.1). Here the speaker does not at some point 'learn' that the referent of the first utterance and the referent of the second utterance are the same person (namely, himself); he might simply know that all along, or never even consider the relevant identity question in the first place. So the speaker never had two distinct files about which we may wonder: "Have they been merged after the relevant identity fact

[31] For discussion see Strawson (1974), Millikan (1997), Recanati (2012, 2016).

was learned?" The response we are discussing thus seems moot with respect to our main cases, no matter whether one accepts or rejects the merge model.[32]

Summing up our discussion in this section, rejecting premise (1) means abandoning a very plausible principle without thereby being able to solve the puzzle. Clearly, a different solution is needed.

6 Rejecting Frege's Constraint?

We initially presented Frege's Constraint as an all-or-nothing thesis: *whenever* a subject rationally holds contrasting attitudes towards a thought t^A and a thought t^B, then t^A and t^B must be distinct. However, some think it's desirable to enable special circumstances in which a thinker's rationality is unaffected by her taking contrasting attitudes to *one and the same thought*. In order for that to be possible, one has to accept the possibility that rational thinkers could be mistaken about the identity conditions of their own thoughts – sometimes, perhaps only in very unfortunate circumstances, taking one thought to be two, or two thoughts to be one.

Now, the possibility that is of interest to the present discussion is the former (rationally taking one thought to be two), not the latter (taking two thoughts to be one). Still, a great part of the discussion of these issues – usually under the banner of debates about the "transparency" of thought –[33] focuses on the second, i.e. cases where, due to unlucky circumstances, a subject treats two concepts that do not even co-refer as being the same. Examples usually include cases of mistakes in perceptually tracking objects. For instance, a subject is tracking a bee flying in her visual field and thinks 'this bee is buzzing'; unbeknownst to her, she starts tracking a second bee as the first swiftly crosses its path; so she thinks 'this bee is [still] buzzing' and then infers 'one and the same bee was buzzing all along', assuming she has successfully kept track of the same bee throughout the perceptual episode. It seems that the premises leading to the conclusion

[32] There are other possible objections against premise (1) of our puzzling argument. For instance, one might argue that Wendy is expressing surrogate metarepresentational thoughts that mimic Peter's confused point of view (see Recanati's discussion of "vicarious files" in chs. 14-15 of Recanati 2012). We think that strategy would only have any plausibility in cases where a thinker knows about the other's confusion, so it would not apply to the first two variants of the puzzle, but we will not develop this here for reasons of space.

[33] The term 'transparency' comes from an influential paper by Boghossian (1994). For a recent survey of the debate, see Wikforss (2015).

do not even refer to the same bee, so that the reasoning is simply invalid. But the inference seems rational – in an intuitive sense, the subject was unlucky in drawing an invalid inference. It has thus been suggested that we simply accept that this subject is rational, despite falsely believing that she is referring to the same bee all along (Gerken, 2011; Schroeter, 2007).

Does this idea have consequences that are relevant for our cases? That is, if we accept that one can rationally take two different thoughts to be one, should we also accept that one can rationally take one thought to be two? It is not clear that we should. One might hold that taking two thoughts to be one can be rational, while taking one thought to be two is always irrational. Indeed, while arguing that in some cases one could rationally treat two non-coreferential thoughts as the same, Schroeter (2007, p. 602, fn. 6) points out that Frege's Constraint is compatible with her view and explicitly rejects the possibility that a rational subject might treat one and the same thought as two.

In any case, there are authors who explicitly argue for the possibility of rationally taking one thought to be two. For instance, this is how Sainsbury & Tye (2012, pp. 131-138) analyze the Kripkean story about Paderewski. Their own account of the case rests on the claim that Peter accepts and rejects what is in fact the very same thought; however, Peter is not irrational, since he thinks (wrongly) that the thought he accepts is distinct from the thought he rejects. Peter's rationality is thus safeguarded, because of his ignorance about the identity of his own thought.

There is no space to go into the deeply ramifying implications of this view. It will be enough for our purposes to invoke, once again, the indexical variant of the Kripkean puzzle. More specifically, we believe that, even if Sainsbury & Tye's story about Kripke's original case works, it does not seem to extend to the indexical variant. It is not hard to see why this is so.

Sainsbury & Tye's interpretation of the Kripkean story is based on their own originalist theory of concepts. According to originalism, two concepts are the same if and only if they have the same origin. We do not need to go into the details of their view to see that, in a very intuitive sense, Peter's two thoughts t^1 and t^2 are constituted by concepts of Paderewski that have the same origin. There is a network of interconnected uses of the single name 'Paderewski' that causally and historically descend from a single original use – for instance, Paderewski's parents "baptizing" the child with that name. In simpler

terms, Peter's concepts derive from two uses of the *same name*. So the two concepts have the same origin; therefore, they are the very same concept. So far, so good.

The problem is that this strategy doesn't generalize – even if we grant that it succeeds with proper names, it fails to apply to indexicals. Consider for instance the indexical variant of Kripke's case that we have previously devised. In that variant, remember, there is no public name that Peter takes to be two. Instead, Peter is interpreting Wendy's two utterances of 'You have musical talent', which crucially involve the second-person pronoun 'you'. Now, the second-person pronoun is clearly not a proper name like 'Paderewski'; it is an indexical, one which is often treated as analogous to demonstratives like 'this', 'that' and 's/he' (in its deictic uses).[34] But, as Sainsbury & Tye (p. 51-53) themselves acknowledge, concepts that correspond to indexical expressions are unlike other concepts in many respects:

> It's a feature of indexical concepts that a speaker can introduce them for himself, independently of other thinkers. This contrasts with public concepts acquired by immersion, like the concept PADEREWSKI. It's not up to individual users to settle anything about the nature or semantics of that concept. (Sainsbury & Tye, p. 52, fn. 15)

One consequence of this asymmetry is that, according to Sainsbury & Tye, the identity conditions of indexical and demonstrative concepts are much more closely connected to their thinker's beliefs and intentions. One can argue that Peter's 'Paderewski'-utterances express the same concept even though he thinks they don't, since the concepts expressed by proper names are "public" and "insulated" from the thinker's idiosyncratic intentions. But this is much harder to defend when indexical and demonstrative expressions are concerned. According to Sainsbury & Tye, we can only determine whether two demonstrative concepts are the same by first inquiring whether the thinker herself takes them to be the same. Consider for instance Perry and Evans's famous case of a subject who thinks he is looking at different ships, when in fact he is looking at one very long ship that is visible through different windows. Sainsbury & Tye say:

[34] For instance, Kaplan (1989) classifies the second-person pronoun as a demonstrative. See also Braun (2017).

> The two uses of the complex demonstrative concept-template THAT SHIP involve distinct specific demonstrative concepts [...]. This can be inferred from the speaker's intentions and reactions. For example, he has no inclination to bring forward the information *was built in Japan* when having the thought associated with the view from the second window. [...] two concepts, concepts originating in distinct acts of concept introduction. (*ibid*., pp. 51-52)

We can now apply Sainsbury & Tye's view to our case. Upon hearing Wendy's two utterances of 'You have musical talent', does Peter entertain the same thought? As we have seen, that depends on Peter's beliefs and intentions: does Peter take himself to be thinking the same thought? In our case, he does not – upon hearing Wendy's second utterance, Peter fails to recognize the referent as the same person that she was talking about earlier, so he takes himself to be thinking a *different* thought about a *different* person. This means that, just like the subject in the ship case, he has "no inclination to bring forward the information" acquired in the first context and apply it to the man in the second context. So Sainsbury & Tye would have to grant that Peter's two concepts are distinct, as they do with the subject in the ship case. In Sainsbury & Tye's terminology, Peter's token concepts have distinct origins – his two independent encounters with the referent – so they are distinct *tout court*.

Now, Sainsbury & Tye could respond that, when Peter interprets Wendy's two utterances, his thoughts somehow have Wendy's thoughts as their origin; but Wendy expresses a single thought through her two utterances (Section 5), so Peter's thoughts have the same origin and count as identical after all. The first problem is that this is simply incompatible with Sainsbury & Tye's own account of indexical concepts. As we have seen, Peter is just like the subject in the ship case: they both fail to recognize the object they're in perceptual contact with, thus failing to "bring forward" the information acquired in their first encounter with the referent. In the ship case, Sainsbury and Tye postulate distinct concepts because of the subject's failure in bringing forward the relevant information, so they would have to do the same in Peter's case. But there is a second, more fundamental problem: Peter's concepts *do not* originate in his exchanges with Wendy. At the party, Peter sees a man and forms a demonstrative concept for him;

then Wendy makes her utterance and Peter identifies the referent as the man in question, someone he is thinking about in a demonstrative manner. At the political rally, Peter sees the man on stage and forms a demonstrative concept, then Wendy makes her utterance and Peter identifies the referent as the man he is thinking about demonstratively. So the origins of Peter's concepts are not Wendy's two utterances, but his two independent sightings of the same man; and surely these are *different* origins by Sainsbury & Tye's own lights.

Summing up, even if Sainsbury & Tye's originalist account was successful in Kripke's case, it would still fail to apply to our indexical variant, where there is no reason to postulate a single origin for Peter's concepts. So it remains much more plausible to hold that Peter's thoughts t^1 and t^2 are distinct, as Frege's Constraint predicts. Like previous attempts, then, rejecting Frege's Constraint fails to solve our puzzle in its full generality.

7 The thought-transfer model

There are two further assumptions in our puzzling argument that have not been examined yet: the thought-transfer model and the transitivity of thought identity. Both principles are necessary for the argument to go through; reject one of them, and the puzzle will disappear. Could that be the solution?

Before answering that question, note that rejecting the transitivity of thought identity is just one way of abandoning the thought-transfer model. Since numerical identity is transitive, we cannot hold that two thoughts are numerically identical without the relation between them being transitive. What one *could* hold is that, when we say that two agents entertain 'the same thought', we do not mean that there is a single thought that they both entertain. But then what *do* we mean? The proponent of this move will need to find an alternative – a relation *R* that holds between the thoughts of different subjects and is weaker than numerical identity. Before discussing some options, note that this has consequences for the thought-transfer model, which claims that communicating subjects express the same thought: if 'same thought' expresses a non-transitive relation *R*, the model requires something weaker than numerical identity. Rejecting transitivity thus turns out to be one way of weakening the thought-transfer model. We will therefore

focus the rest of our discussion on the possibility of rejecting thought-transfer, treating the rejection of transitivity as one variant of this strategy.

Nowadays the thought-transfer model is not as popular as it used to be, having been recently criticized on various fronts. First, the model has been rejected by a number of authors who adopt Lewis's view about *de se* thought:[35] these authors find Lewis's view independently plausible, see that view as incompatible with the thought-transfer model, and conclude that the latter has to go. Another group of views that put pressure on thought-transfer is formed by conceptual/inferential role theories and internalist theories of content[36] – proponents of these views generally recognize their own theory as incompatible with the idea that communicating agents share the same thought. Finally, many psychologists working on concepts think that people's concepts are radically distinct and even diverge within the same subject in slightly different contexts. This should come as no surprise: psychologists generally take concepts to be determined by the bodies of information individuals retrieve in the course of exercising higher-order cognitive capacities, such as categorization, and these bodies of information have been repeatedly shown to be highly variable from individual to individual (and from context to context).[37]

It will be helpful to use an example to illustrate the conflict between the thought-transfer model and one of the views we have just mentioned: Lewis's account of belief content as centered content. The point is clearly illustrated by Weber (2013), who calls the thought-transfer model the "FedEx model" and Lewis's account "the centered belief account":

> A believes that she has Groat's disease and tells B: ''I have Groat's disease''. If everything goes well, we expect B to learn that A (or the speaker) has Groat's disease, while remaining agnostic about whether she herself has the disease. Call this case *Good Groat's*. In the second somewhat bizarre scenario B acquires the

[35] See for example Torre (2010), Ninan (2010), Weber (2013), and the seven essays in the second part of García-Carpintero & Torre (2016).

[36] See for example Block (1993), Prinz (2002), Schneider (2011), Chalmers (2011), Pagin (forthcoming), Valente (2019).

[37] See for instance Machery (2009). Whether the psychologists' notion of a concept is the same as the one we're concerned with is a controversial matter that need not concern us here.

> belief that she herself has Groat's disease, while remaining agnostic about whether A is diseased. Call this situation *Bad Groat's*. Clearly, communication has failed in *Bad Groat's.* However, a centered FedEx model would classify the cases in the opposite way. According to the centered belief account, the content of A's belief is the set of individuals with Groat's disease. [...] As B shares a belief with the same content in *Bad Groat's* but not in *Good Groat's,* the FedEx model is obeyed in *Bad Groat's* but not in *Good Groat's*. This means that the former case gets misclassified as successful and the latter as unsuccessful communication. [...] Something has to give. As it isn't a viable option to simply reverse our classification, we seem forced to give up either the centered belief account or the FedEx model. (Weber 2013, p. 209)

Weber himself ends up rejecting thought-transfer; others have taken the same path. But these objections to the thought-transfer model have a clear limitation: they depend on the specific view underlying the objection, like the Lewisian view underlying Weber's argument. And while such views have certainly been influential, they have also been rejected by some important figures in the debate. For instance, Stalnaker (1981) objects against Lewis's theory of belief content precisely *because* it is incompatible with the possibility of entertaining the same content in communication. So the conflict between the thought-transfer model and the aforementioned views can cut both ways, depending on how plausible one finds the model to be.

To avoid a stalemate, we propose using our puzzle to provide a fresh perspective on the issue. The puzzle we have proposed does not presuppose highly controversial views about the nature of belief – we have presupposed neither Lewisian centered contents, nor conceptual/inferential role semantics, nor semantic internalism, nor the view that concepts are retrievable bodies of information. Instead, we have shown how a puzzle can be generated from a set of premises that are widely accepted and highly plausible; furthermore, our puzzle does not depend on a narrow selection of cases, since it can be generated with different kinds of expressions (names, indexicals) and different kinds of conversational contexts (contexts where the hearer's confusion is known and contexts where it is not).

In light of these considerations, the thought-transfer model now seems to face considerable pressure. The model is not only incompatible with some influential – but controversial – accounts of belief content; it also appears incompatible with a set of basic principles that theorists of different persuasions accept. It thus seems natural, almost inevitable to reject the model. That would block our puzzling argument, but at what price? Do we have a viable alternative to thought-transfer? Many authors think we do. One popular option is to appeal to similarity.[38] Why think that speaker and hearer must share the *very same thought*? Why can't they just have *similar* thoughts in order for communication to be successful?

A full discussion of similarity-based theories would be beyond the scope of this paper, so we will only highlight one obstacle to solving our puzzle through similarity. Clearly, any similarity view has to specify what properties are relevant when evaluating the similarity of two thoughts. Without this specification, any two thoughts will count as sufficiently similar, since similarity is notoriously "cheap": any two thoughts will share an infinite number of properties.[39] If I say 'Paderewski has talent' and you identify the wrong person as the referent, you have surely misunderstood the utterance. Still, there will be an infinite number of properties in common between our respective thoughts, such as: *being entertained by a subject*; *being entertained in a context where A is speaking*; *being entertained in a context where A is speaking and B is listening*; and so on. If each shared property contributes to the similarity of the two thoughts, then we will eventually reach the desired similarity score. Obviously, this is not the desired result, for it would make miscommunication impossible – speaker and hearer would always communicate successfully, since their thoughts would always exhibit the required level of similarity. This is not by itself a fatal objection to similarity, but it does show that more work is needed before we can tell whether thought identity can be replaced by thought similarity. Once the similarity theorist has told us what the desired similarity metric amounts to, we will be in a position to tell whether this is a satisfactory solution to the puzzle. Until then, the appeal to similarity is at most a promising strategy needing further development.

[38] See for instance Harman (1993), Prinz (2002), Pagin (forthcoming).
[39] See Onofri (2016).

Analogous considerations apply to the notion of "type-identity", which is often invoked by theorists such as Fodor (1998, 2008) and Laurence and Margolis (1999, 2007). According to these authors, mental representations can be classified according to a type-token distinction, much like natural language symbols. So two subjects can share the same thought by tokening representations of the same type, instead of having numerically identical thoughts. But the point made above about similarity also applies here: these views ought to provide us with a set of criteria to decide which types count towards communicative success, and which do not. If the criteria are too strict, communication will be too difficult or impossible; if they are too loose, any two subjects will turn out to communicate successfully.[40]

The same applies to another possible solution to our puzzle. As we have seen, premise (1) of our argument claims that Wendy expresses the same thought *t* in her two conversations with Peter. Instead of rejecting the premise as false – a move we already criticized in Section 5 – one could hold that 'same thought' must here be interpreted as expressing a non-transitive relation holding between Wendy's thoughts at different times. It would then be invalid to derive our contradictory conclusion by applying the transitivity of identity, as our argument does. This kind of view has been recently defended by Prosser, who rejects transitivity in both the interpersonal and the intrapersonal case:[41]

> Strictly speaking the relation that we capture by saying that Brown thinks of O under the same MOP [Mode of Presentation] as Jones is not an identity relation but an intransitive transparency relation. But it is still exactly as true to say that I have retained a belief, or share a belief with someone else, as it is to say that I did various things in the past or will do various things in the future. (Prosser 2019, p. 17)

[40] See Aydede (1998), Schneider (2011) for discussion. Similarity and type-identity are not the only candidates. For instance, various theorists have recently defended views where the *relation* between speaker's thought and hearer's thought determines communicative success, without that relation amounting to a form of similarity or type-identity. For reasons of space, it would be impossible to discuss these proposals here; among others, see Fine (2007), Schroeter (2012), Heck (2012) and Recanati (2016).
[41] See Prosser (2019, forthcoming) and Recanati (2016).

> The ‘same MOP’ relation thus fails to be transitive even in the intrapersonal case. (*ibid*., p. 16)

Prosser thinks the problems arising for thought identity parallel those that arise from fission and fusion cases in the personal identity literature. So he suggests a parallel solution: taking ‘same thought’ to express a non-transitive relation, in the same way as various authors have taken ‘same person’ to be non-transitive. If this move is correct, it is more than enough to solve our puzzle – just by rejecting transitivity for thoughts held at different times, the puzzling argument would turn invalid. But is the move correct after all? As with the thought-transfer model, one would need to closely examine the relation that is being proposed as an alternative to numerical identity. As we have seen in Section 5, some cases are best described as ones where the subject *fails* to entertain the *same* thought across time. If the candidate relation we are proposing is too weak, then we might misclassify these cases as ones where the same thought has been successfully retained. Here we cannot discuss how Prosser and others might respond, but a full solution to our puzzle would certainly need to address the issue.

It’s now time to draw some general conclusions from our discussion of the thought-transfer model. Our puzzle shows the model to be inconsistent with a set of claims that are quite plausible, widely endorsed, and very difficult to reject, as the previous sections have shown. So we believe that, among the principles that generate the puzzle, thought-transfer is the least plausible one. Our conclusion is that communication does not involve transferring thoughts from speaker to hearer. At the same time, we have highlighted that this theoretical choice should not be taken lightly. Sacrificing thought-transfer means having to offer an alternative theory of communication, which might be based on similarity, type-identity, or some further notion. Until these alternatives have been fully developed, we will not know whether thought-transfer can really be abandoned. So our conclusion is tentative: rejecting the thought-transfer model is the least costly option, but it still comes at a price.

8 Conclusion

We have presented a puzzle about understanding. Our argument was based on a set of cases where a confused subject is interacting with a non-confused one. The puzzle

derives from the plausible idea that these two subjects would be able to understand each other, plus a set of attractive, widely accepted principles concerning thought and communication. The problem arises for different kinds of expressions – names, indexicals – and for different kinds of speakers – speakers who don't know about the hearer's confusion, speakers who do. It is thus a general problem which requires a general solution – as we have seen, some strategies seem to work with one variant of the puzzle but turn out to be completely implausible with others.

In order to provide a tentative solution to the problem, we have suggested sacrificing the thought-transfer model of communication. As we have seen, this is not a novel claim in itself; what makes the claim interesting is our argument for it. Our puzzle arises from a small set of claims which are widely endorsed in the literature, not from a highly controversial theory such as Lewis's centered-content model or internalism. So the reasons we have offered against thought-transfer ought to have more dialectical force than other more parochial considerations against it.

Yet our solution to the puzzle is tentative, for abandoning thought-transfer does have a cost: our preferred alternative must yield the right predictions about the relevant cases of communication. If none of the alternatives achieves this, then thought-transfer might still be our best option and a different solution to our puzzle will be needed. So there is no easy solution to the problem we have presented; our puzzle about understanding is a real puzzle after all

References

Aydede, Murat. (1998), 'Fodor on Concepts and Frege Puzzles'. *Pacific Philosophical Quarterly*, 79: 289-294.

Boghossian, Paul. (1994), 'The Transparency of Mental Content'. *Philosophical Perspectives* 8: 33-50.

Braun, David, (2017), 'Indexicals', The Stanford Encyclopedia of Philosophy (Summer 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2017/entries/indexicals/>.

Block, Ned. (1993), 'Holism, Hyper-Analyticity and Hyper-Compositionality'. *Mind and Language* 8 (1): 1-26.

Chalmers, David J. (2011), 'Propositions and Attitude Ascriptions: A Fregean Account'. *Noûs* 45 (4): 595-639.

Crimmins, Mark (1992), *Talk About Beliefs*. MIT Press.

Cumming, Samuel (2013), 'From Coordination to Content'. *Philosophers' Imprint* 13.

Egan, Andy (2007), 'Epistemic Modals, Relativism and Assertion'. *Philosophical Studies* 133 (1): 1-22.

Fine, Kit (2007), *Semantic Relationism*. Blackwell.

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

Frege, Gottlob (1918/1956), 'The Thought: A Logical Inquiry'. *Mind* 65 (259): 289-311.

García-Carpintero, Manuel & Torre, Stephan (eds.) (2016), *About Oneself: De Se Thought and Communication*. Oxford University Press.

Gerken, Mikkel (2011), 'Conceptual Equivocation and Warrant by Reasoning'. *Australasian Journal of Philosophy*, 89 (3), 381-400.

Harman, Gilbert (1993), 'Meaning Holism Defended'. *Grazer Philosophische Studien* 46: 163-171.

Heck, Richard (2002), 'Do Demonstratives Have Senses?'. *Philosophers' Imprint* 2: 1-33.

Heck Jr, Richard G. (2012), 'Solving Frege's Puzzle'. *Journal of Philosophy* 109 (1-2): 132-174.

Kaplan, David (1989), 'Demonstratives: An Essay on the Semantics, Logic, Metaphysics and Epistemology of Demonstratives and other Indexicals'. In J. Almog, J. Perry

& H. Wettstein (eds.), *Themes From Kaplan*. Oxford University Press: pp. 481-563.

Kripke, Saul (1979), 'A Puzzle About Belief'. In A. Margalit (ed.), *Meaning and Use*. Reidel.

Margolis, Eric & Laurence, Stephen (eds.) (1999), *Concepts: Core Readings*. MIT Press.

Margolis, Eric & Laurence, Stephen (2007), 'The Ontology of Concepts: Abstract Objects or Mental Representations?'. *Noûs* 41 (4): 561-593.

Loar, Brian (1976), 'The Semantics of Singular Terms'. *Philosophical Studies* 30 (6): 353-377.

Loar, Brian (1988), 'Social Content and Psychological Content'. In R. H. Grimm & D. D. Merrill (eds.), *Contents of Thought*. University of Arizona Press.

Machery, Edouard (2009), *Doing Without Concepts*. Oxford University Press.

Millikan, Ruth (1997), 'Images of Identity: In Search of Modes of Presentation'. *Mind* 106 (423): 499-519.

Ninan, Dilip (2010), 'De Se Attitudes: Ascription and Communication'. *Philosophy Compass* 5 (7): 551-567.

Onofri, Andrea (2016), 'Two Constraints on a Theory of Concepts'. *Dialectica* 70 (1): 3-27.

Onofri, Andrea (2018), 'The Publicity of Thought'. *Philosophical Quarterly* 68 (272): 521-541.

Pagin, Peter (forthcoming), 'When Does Communication Succeed? The Case of General Terms', forthcoming in T. Marques & A. M. Wikforss (eds.) (2020), *Shifting Concepts: The Philosophy and Psychology of Conceptual Variability*. Oxford University Press.

Peacocke, Christopher (1992), *A Study of Concepts*. MIT Press.

Perry, John (1979), 'The Problem of the Essential Indexical'. *Noûs* 13 (1): 3-21.

Perry, John (1980), 'A Problem About Continued Belief'. *Pacific Philosophical Quarterly* 61 (4): 317-332.

Prinz, Jesse J. (2002), *Furnishing the Mind: Concepts and Their Perceptual Basis*. MIT Press.

Prosser, Simon (2019), 'Shared Modes of Presentation'. *Mind and Language* 34 (4): 465-482.

Recanati, François (2012), *Mental Files*. Oxford University Press.

Recanati, François (2016), *Mental Files in Flux*. Oxford University Press.

Sainsbury, R. M. & Tye, Michael (2012), *Seven Puzzles of Thought and How to Solve Them: An Originalist Theory of Concepts*. Oxford University Press.

Schiffer, Stephen (1978), 'The Basis of Reference'. *Erkenntnis* 13: 171-206.

Schneider, Susan (2011), *The Language of Thought: A New Philosophical Direction*. MIT Press.

Schroeter, Laura (2007), 'Illusion of Transparency'. *Australasian Journal of Philosophy* 85 (4): 597-618.

Schroeter, Laura (2012), 'Bootstrapping our Way to Samesaying'. *Synthese* 189 (1): 177-197.

Soames, Scott (2002), *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity*. Oxford University Press.

Stalnaker, Robert C. (1981), 'Indexical Belief'. *Synthese* 49 (1): 129-151.

Strawson, Peter (1974), *Subject and Predicate in Logic and Grammar*. Methuen.

Torre, Stephan (2010), 'Centered Assertion'. *Philosophical Studies* 150: 97-114.

Unnsteinsson, Elmar (2018), 'Referential Intentions: A Response to Buchanan and Peet'. *Australasian Journal of Philosophy* 96 (3): 610-615.

Valente, Matheus (2019), 'Communicating and Disagreeing with Distinct Concepts: A Defense of Semantic Internalism'. *Theoria* 85: 312-336.

Weber, Clas (2013), 'Centered Communication'. *Philosophical Studies* 166 (S1): 205-223.

Wikforss, Åsa (2015), 'The Insignificance of Transparency'. In S. Goldberg (ed.), *Externalism, Self-Knowledge, and Skepticism*. Cambridge University Press.

3 Communicating and Disagreeing with Distinct Concepts: a Defense of Semantic Internalism

Article published in *Theoria,* 85: pp. 312-336, (2019).

Abstract

I suggest a solution to a conflict between semantic internalism – according to which the concepts one expresses are determined by one's use of representations – and publicity – according to which, if two subjects successfully communicate or are in genuine agreement, then they entertain thoughts constituted by the same concepts. My solution rests on the thesis that there can be successful communication and genuine agreement between thinkers employing distinct concepts as long as there is a certain relation (of conceptually guaranteed sameness of extension) between them. In section 2, I motivate semantic internalism and show how it conflicts with publicity. In section 3, I carve the logical space of possible solutions to the conflict into liberal and conservative solutions. Section 4 assesses Wikforss' conservative solution to Burge's arthritis thought-experiment and concludes that it fails for more than one reason. Section 5 introduces a new case study involving a deferential concept. This case serves as the backdrop for my positive account offered in section 6. The conclusion of the paper is preceded by a comparison of my view with another recently proposed by Recanati (section 7) and some replies to possible objections (section 8).

1 Introduction

Roughly, the main aim of this work is arguing for the compatibility between the view that our use of representations determines the concepts we express with the wholly public nature of concepts. More precisely, I will argue that the concepts we express are individuated by our dispositions to apply mental and/or linguistic representations only to certain scenarios and will strive to show how this can be made compatible with the general methodological principle according to which successful communication and/or genuine agreement between subjects require that they share their concepts. Before proceeding, some setup is needed.

Concepts, *qua* theoretical entities in the philosophy of language and mind, are the constituents of our thoughts. They are the building-blocks of the contents of our propositional attitudes, such as beliefs, desires, fears and conjectures. When one believes,

for example, that a bachelor is an unmarried man, one has a belief whose content contains, among others, the concept BACHELOR.[42]

There is much discussion about the ontology of concepts, especially regarding whether they should be treated as token mental representations or as abstract aspects of content expressible by mental and/or linguistic representations. This is not a paper about the ontology of concepts. Indeed, under the assumption that token mental representations can be sorted into types with respect to the property that individuates content, the arguments developed in here are compatible with any of the two approaches. It is, however, difficult to advance substantial theses about concepts while, at the same time, remaining ontologically neutral. It is only for the sake of simplicity that I will treat concepts as aspects of content and will phrase the arguments accordingly. Thus, one of the paper's objectives will be *assessing the conditions for a representation to have expressed a certain concept*. Had I adopted the other approach, I would have talked about the conditions for a token mental representation (i.e. a concept) to be of a certain type. Not much of substance will hinge on this. As a general rule: if one prefers to think of concepts as token mental representations, then one can translate my claims about concepts as being about types of mental representations.

In the next section, I introduce a more contentious claim about concepts, namely, that they are individuated by reference-determining rules.[43] While this is far from a universally accepted thesis, it is a central assumption of the web of views I intend to defend - those which subscribe to the view I call Internalism. After showing that Internalism seems to conflict with the public nature of concepts, I will divide the possible ways out of the conflict into a conservative and a liberal camp (section 3). After siding with the conservatives and pointing out the shortcomings of Wikforss' (2001) own conservative strategy (section 4), I will offer, on the basis of a case study involving deferential concepts (section 5), a positive view according to which subjects can successfully communicate and genuinely agree by holding distinct concepts that are conceptually guaranteed to pick out the same things (section 6). Before concluding (section 9), I compare my view to Recanati's (section 7) and answer to a few objections (section 8).

[42] I will employ the convention of using words in capital letters to refer to concepts.
[43] Thus, one who adopts the view that concepts are token mental representations is invited to think of them as typeable with respect to the reference-determining rules guiding their employment.

2 Internalism and Publicity

This assumption will guide our discussion: a concept's identity-conditions are given by what it would take for an object to fall under its extension. In other words, concepts are individuated by extension-determining rules (henceforth, rules). Thus, e.g. the concept BACHELOR is the concept it is because for something to fall under its extension is for something to be an unmarried man. Alternatively, we can say that BACHELOR is individuated by the rule *that it refers to x iff x is an unmarried man*. Now, it should be obvious that most of our ordinary concepts are not like BACHELOR, whose rule is so sharply expressible and about which everybody would agree. On the contrary, many of the concepts we routinely employ are such that they give us the conflicting feeling of being both competent in their use and unable to explain what they refer to. That feeling notwithstanding, there must be something that makes it the case that some, but not all, applications of a concept are correct, however vague and fugitive that might be.

One should not read me as committed to an outdated picture according to which every concept has a rule expressible as sharp necessary and sufficient conditions. As I said, few concepts are bound to be like BACHELOR. A rule might be significantly complex, e.g. involving relations of family resemblance, relations of typicality, reference to sense-data, motivational states, linguistic tokens, mental tokens, the opinions of experts etc. My talk of rules should then be taken as noncommittal as possible.[44]

We have discussed what makes a concept the concept it is and how that relates to what an object would have to be like for it to fall under its extension. A further question is: what does a thinker have to be like for us to correctly characterize her as having expressed a particular concept by a token representation? Accounts of what it takes to express a concept fall into one of these two camps: internalist or externalist. Internalists claim that which concepts a thinker expresses supervene on matters internal to that thinker, where "internal" should not be read in the sense according to which blood cells are internal, but that according to which one's reasons are. To an internalist, successfully

[44] By assuming that concepts are individuated by reference-determining rules, I approximate my view to a broader research project that includes Jackson (1998), Braddon-Mitchell (2004), Chalmers (2011a) and many others. These authors are often referred to as neo-descriptivists, but this label can be misleading – it gives the wrong idea that concepts are individuated by a purely qualitative description of their extension. None of these authors would agree. Thus, I prefer to use the more neutral term 'rule' and to emphasize that there are, in principle, no limits to what might constitute them.

expressing a concept is a cognitive achievement one is responsible for; we are – to use an expression from Braddon-Mitchell (2004) – the masters of our meanings. There is more than one way to put that idea in the form of a thesis about concepts, but I take the core internalist claim to be that, when one expresses a concept, that is to be explained by that subject being in some kind of personal-level cognitive relation to what makes that concept the concept that it is. More particularly – and drawing on our previous discussion of concepts and rules – I take the core internalist thesis to be:

INTERNALISM: For any concept X and representation Y, a thinker expresses X by Y if and only if this thinker's use of Y is guided by X's rule.

In line with the previous comments, 'using a representation under the guidance of a rule' has to be understood in the sense that the relevant subject, if asked, would, in principle, be able to explain what he meant by coming up with the relevant rule. In reality, things are seldom that straightforward. As previously noticed, we are seldom capable of explaining what we mean – that is, which rule is guiding us – by the representations we produce. But even in those cases, the internalist contends, the subject has some disposition to apply the relevant representation only to certain cases and not to others, and if that subject were presented with a list of actual and imaginary scenarios, we would, in principle, be able to abstract a rule (and thus figure out which concept it is that she is expressing) from considering every case to which she feels disposed to apply the representation and those to which she does not.

Internalism has more or less been the default position in the philosophy of mind and language until very recently.[45] However good its pedigree, it has come under serious attack in the latter half of the 20th century from the so-called externalists.[46] Externalists often emphasize our inability to come up with definitions for the representations we employ and argue that, even when we manage to come up with candidates, they are usually half-baked and too indiscriminate to be of any real use. The particular arguments these philosophers advance usually involve subjects who manage to express a particular

[45] As Johnston & Leslie remark (2012, p. 116), "something like this substantial picture has a good claim to be at the motivating core of what was once called 'analytical philosophy' - from Gottlob Frege, Kurt Gödel, A.J. Ayer, H.P. Grice and Roderick Chisholm through to George Bealer and Frank Jackson".
[46] Among those, Putnam (1975), Burge (1979), and Kripke (1980) stand out.

concept regardless of being mistaken about its rule or even completely ignorant about its nature. Drawing on these cases, externalists usually emphasize the role of e.g. one's social environment in the determination of the concepts we express.

The main aim of that paper is not to adjudicate between Internalism and Externalism. As is often the case with foundational questions in philosophy, the best way to defend a particular view is not by direct argumentation – as if the decision were just a matter of logical deduction – but by showing that it successfully accommodates the theoretical *desiderata* and that it is resilient on the face of criticism. I believe that the contemporary criticism against Internalism is not as convincing as some philosophers make it out to be and will argue that there exist interesting internalist ways out. Let me summarize what I take to be the core externalist criticism of Internalism.

As mentioned, externalists emphasize cases where subjects seem to express concepts by representations whose use is not guided by the appropriate rules. These cases usually come in two varieties. In the first variety we have subjects who are close to being fully competent with a certain concept but are nonetheless mistaken about some of its crucial features. I take Burge's (1979) notorious story about Bert, an individual who thinks he has arthritis in his thigh (even though it is an inflammation that only affects the joints), to be an example of that sort of case. We'll focus on these in section 4. The other variety of cases involves subjects who manage to express a concept by deferring to others, as when one manages to think and talk about mega-bytes (or fascism, or baroque art etc.) without having the faintest idea of what they are. Deferential cases will be the focus of section 5.

The argumentative pull of these two types of cases obviously depends on the assumption that these individuals indeed manage to express the relevant concepts regardless of not being internally connected to their rules in the way Internalism predicts. As will become clear, this assumption is based on the tendency we have of classifying individuals who successfully communicate or who genuinely agree about some subject matter as sharing their concepts. That tendency should not come as a surprise, given that "one of the core explanatory role of concepts is to capture our most basic ways of keeping track of a topic in thought" (Schroeter & Schroeter, 2016, p. 5). It is not unusual to express this idea as a general constraint on a theory of concepts:[47]

[47] The name 'publicity' comes from Onofri (2016).

(PUBLICITY) Whenever two subjects (I) successfully communicate or (II) are in genuine agreement with each other, then that must be accounted for by them sharing a concept. More specifically,

(I) if A successfully communicates to B a thought containing the concept C^1 by means of an utterance U, then B must entertain a thought containing a concept C^1.

(II) if A genuinely (dis)agrees with B with respect to B's utterance U that expresses concept C^1, A must hold a corresponding attitude to a thought containing a concept C^1.[48]

The most vivid way of arguing in favor of Publicity is by considering a story from Loar (1976) whose conclusion is that a speaker and a hearer who entertain co-referential thoughts have not necessarily succeeded in successfully communicating (or in being in genuine agreement):

> Suppose that Smith and Jones are unaware that the man being interviewed on television is someone they see on the train every morning and about whom, in that latter role, they have just been talking. Smith says 'He is a stockbroker', intending to refer to the man on television; Jones takes Smith to be referring to the man on the train. Now Jones, as it happens, has correctly identified Smith's referent, since the man on television is the man on the train; but he has failed to understand Smith's utterance.

Since Loar's characters fail to successfully communicate regardless of holding thoughts that are true in just the same conditions, one concludes that this relation (*mutatis mutandis* for genuine agreement) requires more than a match of referential content. Sameness of concept is then expected to fill that gap.

The two types of cases we will discuss are instances of cases where Internalism pushes us into claiming that the relevant individuals have distinct concepts but, because they are either successfully communicating or in genuine agreement, Publicity pushes us

[48] I will be exclusively concerned with a notion of (dis)agreement with respect to [the thought expressed by] an utterance. This does not mean that there aren't other philosophically interesting cases of (dis)agreement. I'll get back to this on section 8.

in the opposite direction. I take it that most externalist arguments against Internalism are based on its conflict with Publicity.[49] Indeed, even authors who are sympathetic to Internalism admit that, unless some story is told about how people who have distinct perspectives or understanding about a certain subject matter can nonetheless communicate and stand in agreement about it, then it will fail to attract many followers.[50] Thus, if I manage to show how Internalism can be made compatible with Publicity, then the internalist side of the debate will have gained a significant advantage over its adversary.[51]

3 Liberal and Conservative ways out of the conflict

Let us distinguish two types of ways of solving the conflict just presented. The liberal ways out are those that outrightly dissociate concept expression or successful communication (genuine agreement) from rules. The conservative ways, on their turn, are those that weaken, but do not completely eliminate, the explanatory relation between them:

[49] Not only Burge's arthritis case, as we will see in section 3, but also Kripke's (1980) famous semantic argument against descriptivism. The semantic argument contends that we do not associate sufficiently discriminating rules with the names we use and that sometimes we even associate the wrong rules with them. The usual examples are that of a subject who is competent in talking about the physicist Feynman by means of the proper name 'Feynman' even though she knows no more about him than that he is some famous physicist, i.e. the rule she employs does not determine one and only one referent. Another type of example is that of one who associates 'Albert Einstein' with the rule that it refers to the inventor of the atomic bomb. This individual is grossly mistaken but nonetheless seems to succeed in referring to Einstein. Both types of examples can be seen as presenting a conflict between Internalism and Publicity, since the relevant individuals seem to successfully communicate and think about Feynman or Einstein regardless of the faulty rules guiding their uses.

[50] One very illustrative example is Chalmers' (2011a, p. 14, 18) crucial employment of the notion of 'S-appropriateness' to account for true belief ascriptions involving concepts ('primary intensions' in his terminology) that are distinct from the ones expressed by the ascribee. Chalmers admits that he has no satisfactory account of it, but nonetheless gives it central importance in his account. Under the assumption that a theory of belief ascription can be extracted from a theory of successful communication and genuine agreement, the view I will defend in Section 6 under the name of 'Publicity*' can then be seen as complementary to Chalmers' project. In summary, I believe that knowledge of sameness of extension based on concepts' rules could help account for S-appropriateness in Chalmers' context.

[51] The two types of cases which I will focus on, i.e. that of mistaken and of deferential subjects, do not form an exhaustive list. Cases of cognitive dynamics invite a similar conflict between Internalism and Publicity, and so do cases of conceptual stability across theory change, e.g. if one thinks that ATOM is the same concept today as it was for John Dalton in the early 19th century, then it seems that the same concept has survived unscathed through major revisions in the rule which constitutes it. Hopefully what I argue for the simpler cases will be proven relevant to these more complicated ones, although this claim will be left for future investigation.

I. **Liberalism about Internalism:** the concept one expresses by a representation is independent of the rule guiding one's use of it.

II. **Liberalism about Publicity:** a subject A may successfully communicate a thought that contains the concept C^1 to B even if B entertains a thought that contains the concept C^2, where $C^1 \neq C^2$, and where this communicative success is not grounded on any semantic or epistemic properties of the rules associated with C^1 and/or C^2.

III. **Conservativism about Internalism:** the concept one expresses by a representation is *weakly* determined by the rule guiding one's use of it.

IV. **Conservativism about Publicity:** a subject A may successfully communicate a thought containing the concept C^1 to B even if B entertains a thought containing concept C^2, where $C^1 \neq C^2$, but where this communicative success is grounded on some semantic or epistemic property of the rules associated with C^1 and C^2.

Liberalism about Internalism has been the preferred strategy of externalist philosophers, such as Burge (1979). It is equivalent to a rejection of Internalism; thus, one is free to give an alternative account of what it takes to express a concept – perhaps one in which one's community, or at least its experts, determine which concepts one gets to express, even if one has no cognitive relation whatsoever to the rules in the minds of the experts. Liberalism about Publicity entails that communicative success should be accounted for by matters orthogonal to the rules guiding one's uses of representations.[52] It is not my objective in this paper to argue against these two liberal ways out of the conflict. Instead, I will take their "revolutionary" character to entail that they only become real theoretical contenders as soon as the more conservative ways out are out of the game. Since I think there are good conservative ways out there, I will refrain from considering the liberal accounts more in depth.

Conservativism about Internalism maintains the connection between concept expression and rules but weakens the extent to which one determines the other. One way

[52] Unnsteinsson (2018) claims that the failure of communication in the case from Loar previously presented is due to the subjects having a false belief about the target of the conversation. That could be seen as view according to which communicative success is independent of the subject's perspectives on the subject matter at hand. Cumming (2013) argues that these cases can be explained by the absence of a coordinating convention between the subject's representations. Under a plausible interpretation, this also would amount to Liberalism about Publicity.

of fleshing that idea out is by claiming that, contrary to Internalism, we do not need to be guided by a concept C's constitutive rule in order to express it, but merely be guided by a rule which sufficiently approximates it. I think a view pretty much like that can be extracted from Wikforss (2001). I will consider that proposal in the next section and argue that it fails for at least a couple of reasons.

Finally, Conservativism about Publicity maintains the connection between communicative success and rules but does not entail that the former requires identity of the latter. As a first pass, the idea would be that we can count some people as successfully communicating even when they express distinct concepts, as long as the rules that these people are following are related in such-and-such a way. Naturally, the difficult bit here will be finding a suitable relation between thinker's rules such that, even though they lead these thinkers to express distinct concepts, they can nonetheless be said to be successfully communicating (or genuinely agreeing). I will defend a view like this one in section 6.

4 Burge and Wikforss on arthritis and tharthritis

In this section I consider Wikforss' (2001) defense of an internalist view in the face of Burge's (1979) "arthritis thought-experiment". The way I see it, Wikforss tries to advance a conservative solution against Burge's arguments by means of weakening Internalism. I am sympathetic to Wikforss' ambitions but will argue that her account – or at least the kind of account that can be extracted from her discussion – fails for more than one reason.

Burge (1979) tells the story of Bert, a patient who tells his doctor he has arthritis in his thigh. Since arthritis is an inflammation that only affects the joints, the doctor replies: "No, Bert, you do not have arthritis!". The interesting thing about the story is that we feel compelled to treat Bert and the doctor as successfully communicating by means of their uses of 'arthritis' even though each follows a different rule. Internalism compels us to say that, whereas the doctor expresses ARTHRITIS (the concept individuated by the rule that it refers to a type of inflammation of the joints), Bert expresses the distinct concept THARTHRITIS (the concept individuated by the rule that it refers to a type of inflammation of the joints and limbs). However, since we feel so strongly about counting them as successfully communicating (or as genuinely disagreeing), Publicity compels us

to claim that they are expressing the same concept. The conflict could not be more apparent.

Famously, Burge took his thought-experiment to be a *reductio* of Internalism, which he then discarded in favor of a social externalist picture according to which the concepts one expresses are not determined by things inside one's head (such as rules) but by one's social environment and its linguistic conventions. Instead of biting that bullet, Wikforss notices that Burge's argument depends on the tacit claim that arthritis being a type of inflammation of the joints (and not of the limbs) is part of the rule individuating ARTHRITIS and not just a contingent fact about its referent. In other words, the argument presupposes that Bert commits a conceptual (rule) mistake, as opposed to a merely ordinary empirical one. To see the contrast, imagine that Bert were merely mistaken about some unimportant fact about arthritis, e.g. that he believed arthritis is more prevalent in children than adults. If that was the whole story, it would not be easy to get to any conflict between Internalism and Publicity, since it seems that collateral knowledge about some subject matter (e.g. whether arthritis is an old person's disease) does not get to be part of the rule constituting the correspondent concept. Thus, Burge's argument is supposed to work in virtue of the fact that Bert commits a rule-mistake for a concept which we are nonetheless inclined – because of Publicity-related reasons – to interpret him as expressing.

But why, Wikforss goes on to ask, should we concede that Bert's mistake is so grave that he ends up meaning something distinct by 'arthritis' than his doctor? He is, to be sure, mistaken about an important fact about arthritis, i.e. its scope of occurrence; on the other hand, given Burge's own description of the story, Bert is, overall, a competent user of 'arthritis', rarely subjecting it to inappropriate use. Bert knows many substantial facts about arthritis, e.g. he knows it is a disease that can affect the joints and even that it is a type of inflammation. He is also generally able to apply 'arthritis' correctly in many varied cases. Things would surely be different if Bert were like Schbert, who believes that 'arthritis' applies to round green vegetables with fleshy leaves in the shape of a flower (that's an artichoke). Schbert's use of 'arthritis' is so massively out of tune with the public one that he would best be characterized as meaning a completely distinct thing by the term (i.e. ARTICHOKE).

The contrast which Wikforss strives to make is that between one – like Bert – who is a competent user of a word regardless of being mistaken about some important fact concerning its referent and one – like Schbert – whose use of a word is so idiosyncratic that one is best characterized as expressing a distinct concept by the wrong word. Her intention is arguing that Bert's mistake, regardless of being a grave one, is forgivable:

> It may be that the belief that arthritis afflicts the joints only is central to our understanding of arthritis, but what gives Burge the confidence to say that it is so central that giving it up must imply a change in the meaning of the term 'arthritis'? After all, medical terms like 'arthritis' play a complex role in medical theory, and as always with such terms, it seems possible to have a change in certain parts of the theory, including central parts, without any change meaning. (Wikforss, p. 222)

Wikforss then goes on to remark that almost none of the concepts that matter to us are one-criterion concepts, such as BACHELOR, whose identity conditions seem to be so neatly expressible by a one-criterion rule. In reality, most of our important concepts are much more like living organisms, constantly changing and updating themselves as the need comes. Similar stories abound in the medicine and clinical psychology literature. It seems plausible that psychologists these days have the same concept of autism than did their early twentieth-century predecessors, even though the latter, but not the former, used to define autism in psychoanalytical terms, as opposed to the behavioural-physiological terms preferred nowadays (Majeed, forthcoming).

Now, everybody more or less already agrees about those points: concepts are as dynamic as the theories they are a part of. Indeed, many liberal philosophers (e.g. externalists) have used this type of considerations in order to argue that expressing a concept has nothing to do with the rules one is following. However, and I take it that this is one of the lessons that Wikforss wants to emphasize, this line of argument often fails to acknowledge the great degree of continuity that there is between stages of a concept even when it has undergone radical theoretical changes. To put the same point differently: liberals often emphasize how one like Bert is distinct from his doctor without noticing

how much they have in common. That we tend to count Bert as successfully communicating with his doctor but would not do the same had Schbert been in his place is surely evidence that Bert's mistake is not as grave as it looks. The difference between Schbert and Bert is precisely that the latter is overall in agreement with his doctor about when and where to apply 'arthritis', even if he sometimes commits embarrassing mistakes, while Schbert, on the other hand, is just mistaken all over.

As I read her, Wikforss takes these considerations to support a reformulation of Internalism, according to which there is some flexibility on how much one can deviate from a concept's rule and still successfully express it. The underlying idea is that, as long as one still maintains overall agreement with the proper use of a representation, one's eventual mistakes get to be swept under the carpet. Importantly, this proposal would be a weakening, and not a rejection, of Internalism, since expressing a concept would still depend on having the right sort of rule in one's head. Wikforss never goes so far in her paper, but this is the view I think can be extracted from her considerations:

APPROXIMATION INTERNALISM: For any concept X and representation Y, a thinker expresses X by Y if and only if this thinker's use of Y is guided by a rule that *sufficiently approximates* X's rule.

How much approximation is sufficient will probably change from context to context, but the general idea seems clear enough to dispel Burge's main argument: Bert gets to express ARTHRITIS by his use of 'arthritis' because the rule he follows is sufficiently close to the proper one. Unfortunately, Approximation Internalism fails for more than one reason.

Firstly, it does nothing to explain deferential cases where a thinker expresses a concept she knows close to nothing about. In these cases, the thinker is not guided by a rule which is approximately correct, thus, Approximation Internalism does nothing to explain why we still want to ascribe him the relevant concept. A case of that sort will be the focus of the next session. Secondly, Approximation Internalism seems to suffer from an even deeper problem. If the concepts one expresses are those one's rules more closely approximate, then there is no reason why we should interpret Bert as expressing ARTHRITIS instead of THARTHRITIS, since his rule not only approximates both

concepts', it is identical to the second's. I can think of two ways by means of which one could try to amend Approximation Internalism, but none of them are successful.

The first would be claiming that the concepts we express are the *socially shared* concepts which our rules most closely approximate. Then, since THARTHRITIS is not shared among Bert's community, he ends up expressing its closest public neighbor, ARTHRITIS (call this view Social Approximation Internalism, or SAI for short). The most obvious problem with SAI is that it makes it impossible for any of us to ever express the concept THARTHRITIS, since every attempt to associate a representation with the rule that constitutes it would result in us expressing ARTHRITIS. But it surely should be possible for us to express both concepts if we want – we are, after all, the masters of our meanings. Indeed, I take it that we have been doing just that in our discussion everytime we wrote or read 'THARTHRITIS'.

A second possible refinement of Wikforss' account could make use of the property of *naturalness*, the idea being that the concepts we express are always the ones with the *most natural referents* which our rules approximates (call it Natural Approximation Internalism; NAI for short). NAI is a non-starter for more than one reason. One reason is that it is not even plausible that tharthritis is less natural than arthritis, "since diseases are notoriously bad candidates for natural kinds" (Wikforss, p. 226). A deeper reason would be similar to the one we had against SAI: if NAI is true, we would never be able to think and talk about non-natural stuff, i.e. in attempting to think of an object as being grue we would just end up thinking of it as blue.[53] Both SAI and NAI clearly fail.

In summary, there is one limitation and one problem with Wikforss' discussion. It does not account for deferential cases and it seems to lead us to a problematic account of concept expression. I take those failures to significantly weigh against the strategy of weakening Internalism in order to solve the dilemma which is the focus of this paper. In the next section, I will present a deferential case, show that it also gives rise to a conflict between Internalism and Publicity, and argue that we can nicely take care of it by means of weakening Publicity.

[53] Actually, the most natural color closer to grue could be either blue or green. Thus, NAI would not even succeed in determining a concept for that case.

5 Deferential understanding: Neptune and Schneptune

The following story will be our case study:

(LE VERRIER AND BAPTISTE) The French mathematician Le Verrier, in the year of 1846, predicted the existence of a hitherto unknown planet based on mathematical and astronomical findings related to perturbations in the orbit of Uranus. In order to refer to that planet, he named it 'Neptune' - a name expressing the concept NEPTUNE. It is plausibly the case that, at the time of the introduction of that name, the concept NEPTUNE was individuated by the rule that *it refers to whichever astronomical body is causing the perturbations in the orbit of Uranus*. During those days, Le Verrier used to live with his brother Baptiste, who was very much aware of his brother's new obsession with something he was often referring to as 'Neptune'. Baptiste would often make remarks to his friends and family such as "all my brother talks about these days is about Neptune", "Le Verrier does not even leave his lab anymore because he is so concerned with this Neptune" etc. As it turns out, Baptiste had no idea about what Neptune was apart from that it should be some astronomical thing (possibly a planet or a star), and that this was what his brother was constantly talking about. It is plausible that, regardless of being ignorant about Neptune's nature, we should not bracket from accepting that Baptiste and Le Verrier could very well successfully communicate (or genuinely agree) by means of 'Neptune'.

The conflict between Internalism and Publicity should be clear from the way the story unfolds. The rule guiding Baptiste's use of 'Neptune' (let us call it R*) is no more substantial than *the concept expressed by 'Neptune' refers to whichever astronomical body Le Verrier calls by that name*. However, this is not the rule which constitutes NEPTUNE's identity conditions (let us call it R), namely, the rule that *the concept expressed by 'Neptune' refers to the cause of the perturbations in the orbit of Uranus*. Thus, according to Internalism, Baptiste is not expressing the same concept as his brother, but a distinct one constituted by R*: SCHNEPTUNE – a deferential concept dependent on what someone else's representations refer to. On the other hand, we feel that Baptiste and Le Verrier could very well successfully communicate about Neptune. They could,

for example, genuinely agree that Le Verrier's obsession with Neptune is compulsive and needs medical attention. But then, according to Publicity, we must count them as expressing the same concept by means of 'Neptune'. And thus, the conflict reappears.

A promising idea would be that, in all cases where we are tempted to ascribe a concept to a subject who does not conform to Internalism, that subject expresses the relevant concept *via* deference to other people who in fact do conform to it. At a first sight, there are many things that 'deference' could mean in that sort of context. It could be, for example, a tendency to revise one's use of a concept if one notices that it diverges from the use of others. A bit differently, it could be an obligation to consult an expert if one does not know how to classify some tricky borderline case. These types of deference presuppose that the thinker who is doing the deferring has some independent means of applying the relevant concept which does not involve just blatantly mimicking an expert. Bert's case showcases that type of deference: his grasp of ARTHRITIS has some life of its own, so to say.

In contrast to that case, there are cases where one's ability to express some concept is (almost) completely dependent on what the experts do or say. Think of a person who has heard of black holes but who only knows that they are something physicists talk about. It seems that this person's concept BLACK HOLE doesn't have much of a life of its own. It is, however, still useful in a certain minimal sense; imagine a librarian deciding whether to put a book about black holes in the physics or chemistry sections. Thus, the more we know about some subject matter, the less deferential our concept is and the more things we are be able to do with it.

I think it is clear that the Baptiste's concept is of that latter kind, i.e. he does not have many means of applying it unless he is strictly following in Le Verrier's footsteps. That does not, however, make his concept useless from a cognitive-epistemic point of view. Even if he does not have many means of applying it to the world, there are many reflective uses of concepts that he can engage in, such as wondering what Neptune could be or trying to discover more about what it is by asking his brother to tell him more about Neptune. These reflective uses seem to presuppose the ability to express the relevant concept.

Now, how can deference of such a kind enable us to solve the present conflict? As Greenberg (2014) notes, there are three ways in which deference could come in the help of an account of concepts:

1. Deference enables one to express the same concept as the expert does because deferring to an expert enables one to satisfy the same criteria for concept expression as the expert satisfies.
2. Deference enables one to express the same concept as the expert does because deferring to an expert provides a second way for expressing a concept which is not identical to the criteria that the expert satisfies.
3. Deference enables one to express a concept that is distinct from the expert's but is somehow intimately related to it.

It is easy to see that, if we take the original formulation of Internalism, option 1 is an obvious non-starter. For deference to fulfill the role option 1 prescribes it would have to enable someone like Baptiste to, merely in virtue of deferring to Le Verrier, associate R with 'Neptune'. It is clear that this is not the case. Option 2 provides an interesting way out of the problem. From this option, it follows that there is more than one way by means of which one could express a concept. Thus, even if Le Verrier expresses NEPTUNE in virtue of associating 'Neptune' with R, it could very well be that Baptiste expresses the same concept in virtue of satisfying some distinct criterion. That criterion could very well be simply deferring to someone who is able to express the relevant concept. On closer look, however, option 2 is unsatisfactory. Notice that it strives to save Internalism but ends up having to reformulate it as the following disjunctive thesis: one expresses a concept X by representation Y either if one associates Y with X's rule OR if one defers to someone who satisfies the first condition. However, what was most interesting about Internalism was how neatly it accounted for expressing a concept in terms of the personal-level cognitive mechanisms that thinkers employed (i.e. the rules they followed and the explanations they could give of their uses of a representation). This virtue is evidently lost when one adds a proviso to Internalism allowing that, on top of the usual way of expressing a concept, one gets to achieve the same feat by doing something completely different:

> "[…] a proviso that a thinker can have a thought involving a particular concept in virtue of his deferring with respect to the use of the concept or the concept-word is not a minor addendum to a theory committed to the view that to have a thought involving a particular concept is to exercise the concept's canonical disposition [rule]" (Greenberg, 2014, p. 277)

In other words, option 2 does not make Internalism compatible with cases of deferential understanding so much as it tries to sweep the problem under the rug by advancing an *ad hoc* account without independent evidence in its favor.

The failure of the first two options leave us with the last contender. Option 3 bypasses our intuition that people like Baptiste literally express the same concept as the people to whom they defer, however, it promises to explain our disposition to classify them as being able to successfully communicate by pointing to some relation between their concepts which is distinct from identity. Thus, this option entails that there could be successful communication (and genuine agreement) between people who do not express the exact same concepts as long as there is some special relation holding in between the concepts they do in fact express.

6 Conceptually guaranteed sameness of extension

Choosing to go with option 3 means conceding that people employing distinct concepts can nonetheless engage in a successful conversation and stand in a genuine disagreement with each other even when their thoughts contain distinct – but suitably related – concepts. The plausibility of this view of course depends on the relation it characterizes. At least one thing is clear: that relation must be such as to make it obvious to the relevant thinkers that they are not speaking past each other, i.e. to guarantee convergence on one and the same thing in a transparent way.

As we have already seen, sameness of extension is not enough to play that role since people employing co-referential concepts might nonetheless be speaking past each other. That's precisely what we have seen when confronted with Loar's story back in section 2. Let's go over it again. What seems to explain why the subjects in Loar's story are talking past each other – regardless of referring to the same person – is the fact that

the co-reference between their concepts is a matter of luck. Indeed, the concept Smith expresses by means of 'He' and the concept Jones takes him to be expressing are only accidentally co-referential, i.e. were they not unusually lucky, they would have ended up picking out very different things.

Thus, there is some *prima facie* plausibility to the idea that two subjects are not ready to successfully communicate unless their concepts non-accidentally have the same extension, e.g. unless their co-reference is somehow guaranteed. Going back to our deferential story, one could then argue that Baptiste and Le Verrier's uses of 'Neptune' – even though they express distinct concepts – are guaranteed to co-refer in virtue of the concepts expressed and that this is what explains them being able to engage in the relevant interpersonal relations. This is good as a first pass but much more needs to be said.

What exactly does it mean for two concepts to be conceptually guaranteed to co-refer? As a first bet, it seems that two concepts are thus related when the relevant thinkers can know that they co-refer (if both refer at all) *exclusively* on the basis of understanding the rules which constitute them.[54] Here is one model of how that could happen: if it's logically necessary that two distinct rules can only be satisfied, at the same time, by the same object, then anyone who understands these rules can infer that they are guaranteed to co-refer (if they refer at all). Here is a toy example: X is a concept whose rule is *X refers to the one and only F* whereas Y is a concept whose rule is *Y refers to the one and only F-and-G*. Now, it should be clear that one can know, just in virtue of knowing X and Y's application rules, that if these concepts refer at all, then they refer to the same thing. Of course, it is possible that one fails to refer while the other does not, but it is not possible that they refer to distinct things because if the two predicates (F and F-and-G) are uniquely satisfied, then it follows that they are satisfied by the same thing.

[54] Sameness of extension that can be known on the basis of facts that are extrinsic to the concepts' rules or to the representations which express them does not count as conceptually guaranteed co-reference. HESPERUS and PHOSPHORUS can be known to co-refer on the basis of astronomical facts, but not *exclusively* on the basis of their rules (which should somehow be related to the fact that one was observable only in the morning, the other, in the evening). As expected, an old Babylonian who thought that Hesperus was the most beautiful star should not be counted as in genuine disagreement with another who thought the same of Phosphorus. Thus, by 'conceptual guarantee' I mean something closer to apriority than to metaphysical or nomological necessity, although I will refrain from employing this charged notion.

This is the clearest case in which distinct concepts are nonetheless good enough for successful communication (genuine agreement). To see that, notice that we would count a subject who believes X IS ROUND as genuinely agreeing with a subject who believes Y IS ROUND even though it might be reasonable for the first to believe that X IS ROUND while disbelieving (or doubting) that Y IS ROUND (since one may be unsure about whether there's an unique F-and-G).[55]

Let's take stock. The mere conceivability of concepts like X and Y already entails that distinct concepts – such that one could rationally take contrasting attitudes towards thoughts differing only in the substitution of one for the other – could nonetheless be good enough for the interpersonal relations of communication and genuine agreement. Additionally, the previous considerations already show that Publicity, in its initial formulation, is false and needs to be weakened. Successful communication and genuine agreement can indeed be instantiated by people who express distinct concepts – as long as they have the same extension (if they pick out anything at all) as a matter of logical necessity.

This 'rule implication' model might very well help us account for what is going on in cases such as Burge's arthritis thought-experiment. If one thinks that Bert's concept THARTHRITIS is individuated by something like e.g. the rule *that it refers to the one and only type of inflammation of the joints and limbs* and that the doctor's ARTHRITIS is individuated by something like the rule *that it refers to the one and only type of inflammation of the joints*, we reach a situation which is structurally analogous to that presented in the last couple of paragraphs. I do not think this is the only way to account for that case and admit doubting whether it is the best one, but it is a theoretical possibility nonetheless.

More pressing to our present concerns is the realization that the rule implication model does nothing to help us understand Baptiste and Le Verrier's case – that should be obvious given that the rules they follow are completely independent of each other, i.e. grasping both rules does not warrant one to infer, or at least not without additional

[55] The claim that these concepts are distinct is independent from the assumption that concepts are individuated by rules; it can be grounded on the more general principle (sometimes referred to as 'Frege's Constraint', see Recanati, 2016a, p. 11-12) that, if one can rationally take contrasting attitudes towards contents that differ solely in the substitution of one token concept for another, then these concepts are not the same.

information, that the concepts they individuate co-refer (if both refer at all). A different explanation must be given for their case. Fortunately, it is enough to put oneself in Baptiste's shoes to realize that there is something he knows in virtue of the rule he follows which guarantees that he will converge on the same object as his brother. Remember that Baptiste's tokens of 'Neptune' are designed to express a concept whose rule is that it refers to whatever Le Verrier is referring to by his tokens of 'Neptune'. Thus, Baptiste knows something he could express by saying: "for any concept my brother might be expressing by 'arthritis', I know that I will co-refer with it by my own tokens of that word". The moral of the story is that, even if Baptiste is completely ignorant of the concept his brother is expressing, he still manages to hook his own concept onto his brother's tokens and thus conceptually guarantees that he will successfully co-refer with it (if it is referring to anything at all).

Deferential concepts, then, allow their users to guarantee co-reference with the thinkers they defer to regardless of there not being any relation between the deferential concept's rule and that of the concept expressed by the deferred party. In other words, by employing a deferential concept we manage to successfully communicate (and even genuinely agree) with people whose concepts we can be completely ignorant about. It is truly an ingenious representational mechanism in that it allows people coming from very different epistemic standpoints to hook onto the same subject matter.

In summary, I have presented two different cases of thinkers who express distinct concepts, but which are somehow in a position to successfully communicate or genuinely agree. In both of these cases there was something about the concepts these thinkers expressed that allowed them to know, only in virtue of the rules being followed, that they were bound to pick out the same thing(s). In the first case – rule implication – this guarantee was ensured by a direct relation between the relevant concepts' rules. In the other – deferential – case, however, sameness of extension is not guaranteed by the relevant concepts' rules. It is based on the fact that a deferential concept is designed to hook onto the representations used by the deferred thinker. What is common between the two cases is that the thinkers in question have some non-empirical way to know that they are converging on the same things, and thus, not speaking past each other. This leads us to the following reformulation of Publicity:

(PUBLICITY*) Whenever two subjects (I) successfully communicate or (II) are in genuine agreement with each other, then that must be accounted for by them being in a position to know – only in virtue of the rules being followed – that their uses of the relevant representations necessarily have the same extension. More specifically,

(I) if A successfully communicates to B a thought containing the concept C^1 by means of an utterance U, then B must entertain a thought containing a concept C^2 such that B knows – in virtue of the rule she is following – that C^2 necessarily has the same extension as the concept expressed by a corresponding token that is part of U.

(II) if A genuinely (dis)agrees with B with respect to B's utterance U that expresses concept C^1, A must endorse a thought containing a concept C^2 such that A could know – in virtue of the rule she is following – that C^2 necessarily has the same extension as the concept expressed by a corresponding token that is part of U.

7 Deference, memory and risk

Let me unpack Publicity* by comparing it to a recent view advanced by Recanati (2016a, chapter 5). This author is focuses on cases of cognitive dynamics, i.e. those in which thinkers need to update their concepts in order to account for changes in the context, such as when the concept NOW, expressible by 'now is F', becomes, at a later time, a memory concept BACK THEN, expressible by 'back then was F'. Recanati's view is that concepts can be individuated more or less finely depending on one's theoretical ambitions. If one is interested in the cognitive perspective of a thinker, then concepts should be individuated by their rules[56], thus, e.g. NOW comes up distinct from BACK THEN. However, if the philosopher is interested in the dynamic or interpersonal continuities between concepts at different times or across different thinkers, then one individuates concepts more coarsely and gets the desired result that e.g. thinking of a time as present can be the same as episodically remembering it.

Recanati focuses on indexical and demonstrative thoughts while I have focused on deferential concepts, but our resulting views bear similarities, particularly with respect to the idea that uses of representation guided by distinct rules can express concepts which are intimately related. One difference – at first sight, merely terminological – is that,

[56]In Recanati's terminology, rules are epistemically-rewarding relations (see Recanati, 2012).

while Recanati talks of fine-grained and coarse-grained concepts[57], I reserve the word 'concept' for the fine-grained entity, i.e. that individuated by reference-determining rules. Thus, what Recanati calls 'coarse-grained concepts' I prefer to refer to as distinct concepts related by conceptually guaranteed sameness of extension.

Terminological choices are usually a matter of taste (specially with such complicated terms-of-art such as 'concept', whose meanings are constantly up for grabs). However, I think there is at least one point in favor of my terminological choice: I can avoid claiming that a deferential concept (such as Baptiste's SCHNEPTUNE) is, even if only in some coarse-grained sense, identical to that expressed by the deferred thinker (such as Le Verrier's NEPTUNE). As we have seen, it is always a contingent fact that a deferential concept co-refers with the concept expressed by the target deferred thinker. For example, in a nearby possible world where Le Verrier used 'Neptune' as a name for his new pussycat, SCHNEPTUNE (the concept that refers to whatever Le Verrier refers to by means of 'Neptune') would refer to the furry animal while NEPTUNE would naturally still refer to the cause of the perturbations in the orbit of Uranus. This makes me think that it would be too much of a stretch of the notion of concept identity to claim that these are, even if only in a derivative sense, the same concept.[58]

Another difference is that Recanati argues that, in every case where thinkers successfully communicate with distinct concepts, we face the risk that only one of the thinkers is failing to refer (2016a, p. 71-94). We previously saw this possibility with the particular case of the concepts of the unique F and the unique F-and-G. Recanati, however, thinks that this possibility is live in every interpersonal and diachronic case.[59] Thus, communication with distinct concepts ends up sounding like a risky endeavor.

---

[57]Actually, Recanati talks of 'static mental files' and 'dynamic mental files'.

[58] Schroeter & Schroeter (2016, p. 14) make a similar criticism that would seem to affect Recanati's view but not mine. They argue that deference only ensures a contingent link between the deferential and the deferred concepts, and that, for this very reason, one cannot claim that they are, in any sense, the same concept. Since I never make that claim, the criticism simply does not hit my account. However, Recanati might be able to evade it by claiming that he acknowledges two distinct notions of concept identity: a strong one in which a deferential concept is distinct from the deferred one, and a weaker one, according to which they are the same. He could then claim that Schroeter & Schroeter's criticism only makes sense if 'the same concept' is read in the strong sense. Whether this is a satisfactory answer is a question that I will leave for future work.

[59] Recanati admits of only one possible type of exception but relegates it to a footnote (2016a, p. 94, ff. 14): "I can look at an old photograph of Paris and think: *'Streets were crowded then',* without having the faintest idea when the photograph was taken." Although Recanati does not go on to discuss these cases, they seem *prima facie* related to deferential scenarios.

As risky as it might really be, it is interesting to notice that the possibility of one concept referring while the other does not is not live for cases where co-reference is guaranteed by means of a deferential concept. There just is no possibility that, e.g. Baptiste and Le Verrier are in a situation where only one of them is failing to pick out a proper referent. It is slightly ironic that deferential concepts, although being a product of thinkers in impoverished epistemic situations, allow no possibility of failure similar to that of "rule-implicated concepts". The referential success of a deferential concept depends exclusively on the deferred concept's.

I once believed that memory was another (the only other) type of thought that possessed that same property. When a perceptual concept – those we usually express by a demonstrative when perceiving an object – becomes, at a later time, a memory concept – those which we usually express by a demonstrative when recollecting –, it seems conceptually impossible that only one of them fails to refer. If that were right, the conclusion would be that memory and deference are privileged forms of thought in at least this one aspect, regardless of their very different functions and etiology. One could, at this point, even toy with the idea that memory is a form of perceptual deference, in the sense that a mnemonic concept would refer to *whatever was referred to by the originating perceptual concept*. However, Recanati (2016a, p. 89-94) argues convincingly that memories additionally locate the source of their originating perceptual experience in thinker's past (i.e. and not on somebody else's). This, summed up with the possibility of quasi-memories, is enough to entail that a mnemonic concept could fail to refer, while the perceptual experience on its causal origin did not.[60] The cogency of Recanati's argument – as well as the similarities between memory and deference – will have to be examined at some other time. For now, the lesson should be that deferential concepts afford thinkers a degree of confidence in referential match with their peers that possibly no other type of concept does.

[60] As Recanati (2016a, p. 93) comments, quasi-memories were introduced in the philosophical literature by Shoemaker (1970). These come in at least two types: (i) a memory from a subject neurosurgically implanted in the brain of another; (ii) an apparent memory unconsciously fabricated after listening to someone vividly recount their experience. Thanks to an anonymous referee for introducing me to these points.

8 Objections and replies

Thus far, my discussion has focused on singular concepts, but the general lessons reached should apply across the board. However, general concepts seem to bring complications that so far have not been examined. In this section, I examine the possibility of agreement with expressions that do not necessarily have the same extension, and of disagreement with expressions that pick out distinct things. First stop: agreement with context-sensitive expressions.

One could think that Publicity* is incompatible with the fact that we often count people as agreeing with respect to utterances containing context-sensitive expressions even when, given some contextual differences, their expressions apply to distinct things. As Cappelen (2018, p. 107-121) puts it, "we can talk about the same topic even when we change extension". Take the case of 'tall', for instance: there are cases in which we would count two speakers A and B as saying the same thing by 'Rachmaninoff is tall' – and thus as agreeing on what is said – even if, given their distinct contextual stipulations, their 'tall' tokens apply to distinct people (e.g. according to A, people above 1,80m count as tall; according to B, people above 1,90m). The concepts A and B express by 'tall' do not necessarily have the same extension. Indeed, if Rachmaninoff's height were 1,85m, then only one of the utterances would express a true content – doesn't, then, Publicity* entail that they are not in genuine agreement with respect to these utterances? Yes – but that shouldn't be a problem.

There are many types of agreement and disagreement. I have thus far reserved the term 'genuine' to those in which subjects express concepts whose rules somehow guarantee that they have the same extension. The case of 'tall' is, of course, one in which we have the intuition that the subjects are in agreement but where the concepts they express could pick out distinct things. But that just means that one has to account for our intuition without recourse to the literal contents that they express – it is, after all, overly optimistic to expect that genuine agreement with context-sensitive expressions, such as 'tall', will be accounted for in exactly the same manner as with the others.[61] One could say, for example, that our intuition is based on the contingent fact that Rachmaninoff

[61] Given how much has been written on the special character of indexical concepts, especially regarding how hard it is to characterize sameness of thought with them, this move is not at all implausible (see e.g. Ninan 2016 and Valente 2018).

satisfies both A and B's threshold for tallness. Alternatively, one could say that it is based on the fact that A and B's uses of 'tall' follow the same context-insensitive rule: *that it applies to people whose height is greater than some contextually-determined threshold*.[62] Each strategy will have virtues and defects that I won't get into, but these sketches should at least show that plausible accounts of our intuitions of agreement with context-sensitive expressions could still invoke the rules that subjects follow, and thus, be taken as complementary to, instead of against the spirit of, Publicity*.

Other tricky cases involve disagreement with concepts that do not pick out the same things. If C says that Pluto is not a planet, because C thinks that something is a planet only if it clears its neighborhood of other objects, and D disagrees, shouldn't we characterize C and D as being in genuinely disagreement regardless of the fact that the concepts they express by 'planet' are not only distinct but also pick out distinct things? The reply here, as in the previous case, is that the type of disagreement between C and D doesn't need to be classified as genuine. Indeed, as Chalmers (2011b, p. 542) says, "the manifestly verbal dispute among astronomers about whether Pluto is a planet is best understood as a debate in the ethics of terminology". In other words, we can characterize C and D as engaged in a metalinguistic negotiation about which concept to express with 'planet' (Plunkett & Sundell, 2013), and not as in genuine disagreement with respect to the contents they express.[63]

Is the conclusion then that all instances of purported disagreement involving expressions that pick out distinct things should be characterized as metalinguistic negotiations? There is more than one reason for why the answer should be no. For just one, consider the following case.[64] In a recent experiment, half of the Dutch participants

[62] This would be a conception of (dis)agreement that does not depend on sameness of extension, but only on sameness of rule. Such a conception would allow one to explain e.g. how Oscar and Twin-Oscar somehow agree with each other with respect to their utterances of 'water quenches thirst' regardless of referring to different stuff (Putnam, 1975). It also promises an account of the sense in which two subjects who think of themselves by means of the first-personal pronoun somehow think of themselves in the same way. One wonders whether this conception is more fundamental than the one emanating from Publicity*, but since their difference only manifests in relation to rules that are somehow context-dependent, I will avoid that complication.

[63] Plunkett & Sundell (2013) use 'genuine disagreement' to mean disagreements that are, in my terms, genuine, but also significant types of metalinguistic negotiations. I, on the other hand, reserve the term 'genuine' to what they call 'canonical disagreements'. It goes without saying that our disagreement with respect to these issues is merely terminological.

[64] Another reason why metalinguistic negotiations cannot be the whole story is that they might not be able to capture what is at stake in persistent normative and evaluative disagreements, see Marques (2017). I will not touch upon these issues.

who were asked to color the part of a drawing of a human body corresponding to the arm (in Dutch, 'arm'), colored the drawing from the shoulder to the wrist, while the other half colored it to the fingertips (Majid, 2010). Taking the sample as representative of the whole population, should we then conclude that half of Dutch speakers cannot engage in genuine disagreement with respect to utterances containing 'arm' with the other half? This seems extreme. There is no space to work out a full response to this worry, but a promising way out would involve working out a notion of relevancy, such that we could count subjects as being in genuine disagreement if these subjects could know – in virtue of the rules they follow – that the extension of their concepts is the same for all *relevant* possibilities (as opposed to all metaphysically possible ones). The next step in the argument would then be explaining why the divergence between Dutch speakers is not relevant in the context of genuine disagreements expressible with 'arm'.[65]

In any case, I agree with Plunkett & Sundell (2013) that not all instances of substantial disagreement require us to ascribe the same concepts to the relevant subjects. While these authors focus on cases where the disagreement is accounted as a metalinguistic negotiation, Publicity*, if true, entails that others can be accounted by the presence of concepts which are distinct, but nonetheless guaranteed to co-refer or to pick out the same things in virtue of their rules.

The previous objections implied that Publicity* makes it too hard for people to genuinely (dis)agree with each other. A final objection is that Publicity* might instead make it too easy. Notice how easily one can create a concept that is guaranteed to co-refer with someone else's use of a representation: Dolores is traveling in a foreign country whose native language she knows nothing about; she can, however, at every occasion in which she overhears some local produce a sound, create a concept intended to refer to whatever that sound refers to. Dolores can create that concept even if she has no idea whether the sound produced by the local subject corresponded to a whole sentence, a single word, a meaningless hum or an involuntary yawn. If, by sheer luck, it corresponded to a word, then, Dolores' concept is conceptually guaranteed to have the same extension as the concept expressed by the local. Indeed, her situation would be analogous to that of Baptiste and Le Verrier. But that just means that, according to

[65] Pagin (forthcoming) discusses Majid's (2010) experiment and, in response, develops a view along those lines.

Publicity*, her concept would be such that she could be in genuine agreement with the local with respect to utterances containing it, or even able to successfully communicate by it. That's not a desirable consequence; it seems undeniable that Dolores' metalinguistic trick shouldn't allow her to go that far.

This shows that we need a principled way to distinguish cases where a subject's deferential concept allows her to communicate and genuinely agree with the ones to whom she defers (Baptiste's), and cases in which it does not (Dolores'). The crucial difference seems to be that Baptiste's implicit knowledge about Le Verrier's context and communicative intentions allowed him to infer that 'Neptune' is a singular expression, or even that it was related to astronomy. Dolores has absolutely no knowledge about the local's intentions apart from the sounds coming out of her mouth. That seems to be on the right track. How much information one needs about one's interlocutors before one is able to create a successful deferential concept? That interesting question will have to be left for another time.

9 Conclusion

After considering a few pertinent objections and sketching possible replies, my conclusion is that Publicity* promises Internalist philosophers an account of successful communication and genuine (dis)agreement that overcomes the counterexamples often offered on behalf of externalist philosophers. My main thesis is then that the conjunction of Publicity* with Internalism yields an account of concept expression, communication and genuine agreement that is able to endure classical externalist attacks.

Naturally, I have left many questions untouched, e.g. how to account for interpersonal relationships involving context-sensitive expressions, how to characterize the constraints that a thinker must satisfy in order to create a proper deferential concept etc. Furthermore, I have not offered more than indications of how Publicity* would help us with diachronic cases involving the same thinker at different times – the analogy of deference and memory seems like a particularly promising link to investigate. In any case, I rest more than content if the arguments developed in here help views like Internalism to gain momentum.

References

Braddon-Mitchell, D. (2004) Masters of our meanings" *Philosophical Studies* 118 (1-2):133-52.

Burge, T. (1979) Individualism and the mental. *Midwest Studies in Philosophy,* 4 (1):73-122, 1979.

Cappelen, H. (2018) *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Chalmers, David J. (2011a) Propositions and Attitude Ascriptions: A Fregean Account. *Noûs* 45 (4):595-639.

Chalmers, David J. (2011b) Verbal Disputes. *Philosophical Review* 120 (4):515-566.

Cumming, S. (2013) From Coordination to Content. *Philosophers' Imprint* 13 (4):1 – 17.

Greenberg, M. (2015) Troubles for Content I. In Alexis Burgess and Brett Sherman (eds.), *Metasemantics: New Essays on the Foundations of Meaning.* Oxford University Press.

Jackson, F. (1998) *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford University Press.

Johnston, M & Leslie, S. (2012) Concepts, analysis, generics and the Canberra plan". *Philosophical Perspectives* 26 (1):113-171.

Loar, B. (1976) The semantics of singular terms. *Philosophical Studies* 30 (6):353-377.

Majeed, R. (forthcoming) Why the Canberra plan won't help you do serious metaphysics, *Synthese*.

Majid, A. (2010) Words for Body Parts. In: *Words and The Mind. How Words Capture Human Experience*. Ed. by Barbara C. Malt and Philip Wolff. Oxford University Press. Chap. 3, pp. 58 – 71.

Marques, T. (2017) What metalinguistic negotiations can't do. *Phenomenology and Mind* (12):40-48.

Ninan, D. (2016) What is the Problem of De Se Attitudes? In *About Oneself: De Se Thought and Communication*, edited by Stephan Torre and Manuel García-Carpintero, 86–120. Oxford: Oxford University Press.

Onofri, A. (2016) Two Constraints on a Theory of Concepts. *Dialectica* 70 (1):3-27.

Pagin, P. (forthcoming) When does communication succeed? The case of general terms, forthcoming in Teresa Marques and Åsa Wikforss (eds.), *Shifting Concepts*, Oxford University Press.

Plunkett, D. & Sundell, T. (2013) Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint* 13.

Putnam, H. (1975) The Meaning of 'Meaning'. In his *Mind, Language and Reality: Philosophical Papers*, vol. 2, 215-271. Cambridge University Press.

Recanati, F. (2012) *Mental Files*. Oxford University Press.

Recanati, F. (2016) *Mental Files in Flux*. Oxford University Press.

Schroeter, L & Schroeter, F. (2016) Semantic Deference versus Semantic Coordination. *American Philosophical Quarterly*, n. 53 i.2, p. 193.

Shoemaker, S. (1970) Persons and Their Pasts. American Philosophical Quarterly 7: 269-85.

Unnsteinsson, E. (2018) Referential Intentions: A Response to Buchanan and Peet. *Australasian Journal of Philosophy* 96 (3):610-615.

Valente, M. (2018) What is special about indexical attitudes?, *Inquiry*, 61:7, 692-712.

Wikforss, A. (2001) Social externalism and conceptual errors. *Philosophical Quarterly* 51 (203):217-31.

4 On successful communication, intentions and false beliefs

Article published in *Theoria,* online first doi:10.1111/theo.12186, (2020).

Abstract

I discuss a criterion for successful communication between a speaker and a hearer put forward by Buchanan (2014) according to which there is communicative success only if the hearer entertains, as a result of interpreting the speaker's utterance, a thought that has the same truth conditions as the thought asserted by the speaker and, furthermore, does so in virtue of recognizing the speaker's communicative intentions. I argue, against Buchanan, that the data on which it is based is compatible with a view involving Fregean modes of presentation. In the second part of the paper I critically discuss Unnsteinsson's claim that communicative success depends on the absence of contextually salient false distinctness beliefs about the subject matter of the conversation. I argue that this thesis leads to clearly counterintuitive consequences and that no fundamental role must be given to the presence or absence of false distinctness beliefs in one's account of successful communication. The upshot is that we should stick with Buchanan's criterion. I conclude by employing Strawson and Recanati's concepts of linking and merging to show how the criterion I favor is compatible with the fact that, when subjects hold no relevant false distinctness beliefs, communicative success does not seem to be disrupted by the hearer seemingly failing to recognize the speaker's intentions.

1 Introduction

As a first pass, a communicative event is any that essentially involves the transmission of information from a source to a receiver by means of a perceptible signal. This broad conception of communication might be useful for one or another purpose, but it won't lend itself to much philosophical analysis. The first restriction we shall make is to concern ourselves exclusively with linguistic communicative events between a speaker and a hearer, where a speaker asserts some thought by means of a representation (e.g. an utterance, an inscription etc.) which is then supposed to be interpreted – successfully or not – by the hearer. The philosopher of communication then wonders about what the conditions on the hearer's interpretative process are for it to have been a successful one, that is, for that event to have been one of successful communication. Even at that narrower level of abstraction, one cannot help but realize that linguistic communicative events of that type can be evaluated according to distinct and independent standards of communicative success which do not always output the same results. Take, for example, an utterance of the following sentence:

(1) That man [pointing to a photograph of Nietzsche] is the greatest philosopher who ever lived!

There are at least three different ways in which an interpretation of an utterance of (1) could fail. Firstly, a hearer could fail to assign the same standing meanings to the utterance as the speaker does. So, for example, if the confused hearer thinks 'philosopher' means *musician,* while the speaker knows that it just means *philosopher*, the first will have failed to understand what the second meant and, thus, their communicative event would not have been successful. Secondly, a hearer could fail to grasp *what was said* by the utterance even after correctly assigning its words with the right standing meanings. That is most obviously the case when context-sensitive expressions are present. Thus, the hearer could understand that 'that man' is a complex demonstrative expression usually accompanied by some demonstrative act but mistakenly believe that the speaker is pointing to a person standing across the room from them, and not to the photograph. Failure to understand the speaker's demonstrative intentions would also compromise the communicative event's success, regardless of the correct assignment of standing meaning (character). Finally, the hearer could get everything right at the level of standing meaning and what was said, but still fail to draw an important implicature intended by the speaker. That would be the case, for example, if the hearer were oblivious to the fact that (1) was meant as an ironic utterance, possibly in a context where Nietzsche is having his philosophical reputation harshly scrutinized.

Not only do we seem to have at least three different notions of communicative success corresponding, respectively, to correct grasp of standing meaning, what is said and conversational implicatures, these levels' success conditions seem to be thoroughly independent of each other. Pagin (forthcoming) suggests the following communicative exchange as an instance of success with respect to correct grasp of conversational implicature, but failure with respect to standing meaning and what is said:

(2) Anna: Would you like to go to the movies?
Bob: I have to go to her [pointing at Claire] bank.

Anna would be able to draw the implicature that Bob is too busy to go to the movies regardless of whether she thinks Bob is talking about a financial institution or the riverside and regardless of whether she thinks he is pointing at Claire or the person standing right beside her. In other words, (2) would be an instance of communicative success with respect to the implicature that Bob is busy even in a case where Anna fails to grasp what is said by Bob or even the meaning of one of the words of the utterance.

That all being said, if we intend to investigate communicative events as regards the criteria of their success, the first thing to do is to restrict our focus. In this paper, we will only be concerned with success with regards to standing meaning and what is said.[66] Additionally, our focus will be almost exclusively on communicative events employing paradigmatic types of referential expressions: proper names, demonstratives and indexicals. Some of what I will say is supposed to apply for linguistic communication in general, but making sure that all of our examples contain similar types of expressions will help us get a clearer sense on the notion of communication that is at stake. Finally, we will limit ourselves to assertoric exchanges, i.e. those involving properly assertoric uses of sentences in the declarative mood.

Thus, from this point onward, 'communication' will be used to refer to the more specific type of interaction between a speaker and a hearer with respect to the meaning of and/or what is said by an utterance containing a referential expression produced by the former. Analogously, by 'communicative success' I will mean success with respect to grasping what was said by the speaker and/or assigning the relevant utterance with the appropriate meanings. That notion of communication (and of communicative success) might not *perfectly* map onto all ordinary-language uses of the words 'communication' and 'understanding', but I intend it to approximate one of their central and most important ones. Indeed, the main objective of this paper can be described as one of trying to shed light on and trimming the edges of our folk-notions of communication and understanding.

2 Communicative success and identity of truth-conditions

Criteria for communicative success should minimally spell out necessary conditions for a speaker to have successfully communicated a thought to a hearer by means of an

[66] These would seem to boil down to the same thing for *prima facie* non-indexical expressions such as names.

utterance. It is a platitude that one of the most general aims of communication is enabling a reliable transmission of information between thinkers, thus, it is plausible that an instance of successful communication should at the very minimum guarantee that the belief asserted by the speaker and that subsequently entertained by the hearer necessarily have the same truth-value. In other words, the simplest criterion of successful communication is one according to which success is determined by the hearer acquiring a thought with the same truth-conditions as the one the speaker intends to convey; let us call it the *identity of truth-conditions criterion (C1):*

(C1) There is successful communication between speaker S and hearer H iff H entertains, as a result of interpreting S's utterance, a thought that has the same truth-conditions as the thought asserted by S's utterance.[67]

C1 is as desirable as it is simple. If adequate, it would allow us to account for communicative success without having to resort to any theoretically loaded notions. This very simple view is, unfortunately, subject to devastating counter-examples. One can easily conceive of cases where a hearer ends up entertaining a thought with the same truth-conditions as that asserted by the speaker but where, intuitively, communication has not been successful. Let us call these cases 'Loar-cases', owing to their inspiration in Loar (1976). In Loar-cases, it seems that the failure of communication has got to do with the fact that the sameness of truth-conditions between speaker and hearer's thoughts is not generated by the usual process of correct interpretation.

2.1 Modes of presentation and communicative intentions: Buchanan against Loar

Here is Loar's (1976) original story:

(Loar-case 1): Suppose that Smith and Jones are unaware that the man being interviewed on television is someone they see on the train every morning and about whom, in that latter role, they have just been talking. Smith says, 'That man is a stockbroker', intending

[67] Let's ignore the issue of necessary propositions for simplicity's sake. However, if one really wants to take care of those, then the principle could be restated as "there is successful communication between speaker S, who produces an utterance ascribing property F to object a, and hearer H iff H entertains, as a result of interpreting S's utterance, a thought that ascribes property F to object a".

to refer to the man on television; Jones takes Smith to be referring to the man on the train. Now Jones, as it happens, has correctly identified Smith's referent, since the man on television is the man on the train; but he has failed to understand Smith's utterance.

In Loar-case 1, Jones interprets Smith's demonstrative use of 'he' as if it were anaphoric on previous uses made in the conversation, thus leading us to the intuitive conclusion that they fail to successfully communicate regardless of the sameness of truth-conditions between their thoughts. Loar notoriously took these cases to present a refutation of direct reference views of singular thought, according to which the thought asserted by an affirmative utterance of '*a is F'* is exhausted by the singular proposition that *a is F.* Against that view, Loar argued that, on top of the singular proposition asserted by a speaker in a communicative event, we need to take into account *the modes of presentation* (henceforth, MOP) by means of which that speaker conceives that proposition. By introducing MOPs as components of what is meant by an utterance over and above what they refer to, Loar naturally reached the view that successful communication requires, not only identity of truth-conditions between speaker and hearer's thoughts, but identity of MOP as well.

As Buchanan (2014, p. 57) remarks, Loar's argument in favor of a MOP-based view of thought and communication depends on the unspoken assumption that understanding an utterance 'is simply a matter of recognizing *what* the speaker asserted'. If one assumes that and agrees that Jones recognized the exact same singular proposition semantically associated with Smith's utterance, then it naturally follows that Smith's utterance must have semantically expressed some other content.

It does not take much to see that this line of argumentation does not stand on its own feet. Indeed, it is strikingly similar to the influential Fregean-inspired (and by now generally agreed to be faulty) argument in favor of Senses according to which differences in cognitive value (e.g. co-referential identity statements with different informative potential) must be accounted for by the postulation of fine-grained semantic values that determine the reference of a representation. The point is not that those arguments are invalid; the problem is the presumption that their conclusion is inevitable. Indeed, a significant part of the philosophy of language literature of the last century can be taken as going back and forth over the point that there are alternative explanations to the

Fregean *data* which stick to the view that the content of a singular utterance is exhausted by a singular proposition.[68] The main difference between the classical Fregean argument and Loar's is that the latter does not seem to depend on the informativity of identity statements but merely draws upon our intuitions about communicative success – still, they're analogous in fixing the insufficiency of referential content by means of the postulation of a finer type of content.

In the present paper, I am interested in the notion of communicative success in its own right, and not only in the prospects of using it to motivate a general account of thought. That being said, it is of utmost importance to look for the simplest and most neutral way of cashing out the lessons of Loar-case 1, and Buchanan's view might very well be a good place to start.

Buchanan's main point is that we can account for what goes wrong in Loar-case 1 by appealing to an independently motivated account of the role of communicative intentions in conversations – one that is so plausible that everybody is more or less obligated to accept anyways – and thus completely bypass the idea that the content of thoughts and assertions goes beyond what they refer to. The idea, in a nutshell, is that "the kind of misunderstanding that Loar has called to our attention shows that there is *some aspect* of the speaker's communicative intentions that her hearer is failing to recognize" (Buchanan, 2014, p. 64). In other words, Buchanan is suggesting that the failure of Smith and Jones' communication can be accounted by the fact that, even though Jones reached a thought with the same truth-conditions as Smith's, he did not do so by means of properly recognizing Smith's communicative intention to refer to the man in the television. On that view, successful communication requires that the hearer token the same singular proposition as the speaker *as a result of recognizing the proper inferential process intended by the speaker*:

(C2) If there is successful communication between speaker S and hearer H, then H entertains, as a result of interpreting S's utterance, a thought that has the same truth-

[68] Salmon (1986) is a good place to begin surveying the literature on Frege's puzzle.

conditions as the thought expressed by S's utterance and, additionally, H does so in virtue of having recognized S's relevant communicative intentions.[69]

Buchanan's argument is as simple as it is convincing. Surely any theorist recognizes the role that speaker-intentions and their recognition by hearers have in the proper functioning of communication. Even the most die-hard Fregean philosopher (who believes that interpretations are a matter of assigning MOPs to symbols) has to tell a story about how people come up with a particular MOP during an interpretative process, and that story will most likely involve the recognition of the communicative intentions of the source to be interpreted. Thus - and this is where I am in complete agreement with Buchanan - a view that accounts for communicative success by means of the recognition of the speaker's communicative intentions makes the postulation of MOPs redundant.

Now, even though I agree with Buchanan's argument against Loar's conclusions, I think its impact should not be overplayed. It is important not to forget that the target of Buchanan's criticisms is a very specific type of Fregean account of communication, that is, an account which takes MOPs to be constitutive of what is meant by singular utterances. On the other hand, many recent philosophers defending views which they consider to be Fregean-inspired hold a much more deflationary attitude towards MOPs and instead assume that a MOP can be anything as long as it is able to play a set of interrelated semantic/epistemic roles even if, at the end of the day, they do not end up being the kinds of things which are expressed by our assertions nor the building-blocks of our thoughts.[70] Thus, if one is not careful, the difference between a MOP-based view, deflationarily construed, and the type of view which Buchanan wants to defend could

[69] This is a simplified version of the principle that Buchanan (2014, p. 63-64) in fact goes on to defend. His principle includes the concept of an *ib-feature*, i.e. a feature of the speaker's utterance which she intends that the hearer use as a basis for her interpretation. This detail is not relevant for the concerns of the present paper. It should also be noted that C2 only states necessary, but not sufficient, conditions for successful communication. This is in line with how Buchanan himself frames his own discussion, although Bach and Harnish (1979, chapter 5) can be said to have defended the same principle in its biconditional form. The weaker version is good enough for our present purposes: it's the necessity of the principle, not the sufficiency, that will be put into question in the next section.

[70] One recent example is García-Carpintero's (2016) Fregean-inspired presuppositional view, according to which MOPs are like presuppositions in that they are conveyed (presupposed), but not directly expressed, by means of our assertions and thoughts. In his most recent work, Recanati (2016, p. 145-146) also adopts a kind of presuppositional account of MOPs according to which the role that modes of presentation are supposed to play are executed by the vehicles of thought – which he calls 'mental files' - and presuppositional content they convey depending on the type of epistemically-rewarding relations they are based on.

boil down to a few distinct terminological choices. To evade these complications, let us continue to use 'MOPs' with its more restricted meaning, i.e. as contents of thoughts and assertions that outstrip what they refer to.

That being said, Buchanan does indeed present a strong case against MOP-based accounts of successful communication. His argument, however, is not a direct one but one based on the greater simplicity of another available account. It is thus important to keep in mind that a MOP-based explanation of communicative success is still a coherent option on the table, even if it's not the most parsimonious one around. Buchanan, however, also intended to oppose that weaker compatibility claim – and that is where our disagreement lies.

2.2 Misinterpreting a drawing: against Buchanan

On top of arguing that MOPs are not needed to account for Loar-cases, Buchanan (2014, p. 62) claims that the MOP-based view of successful communication cannot account for certain cases which are both (i) analogous to Loar-case 1 in the relevant respects and (ii) suitably explainable by his preferred account. If correct, this could very well constitute a direct argument against any criteria of successful communication based on the postulation of fine-grained semantic contents. I do, however, believe that Buchanan's argument fails and can be adequately answered. This is the case the author comes up with:

(Buchanan-case) In observance of a religious holiday, Smith is forbidden to read, write, or speak for the day. Because Smith is looking so bored, his friend, Jones, tells Smith he will take him to a movie, but they need to decide what to see. It is mutual knowledge between them that a cowboy movie entitled 'Flat-top Mountain' is one of the many movies playing at their local Cineplex. Smith grabs his notebook and draws a mountain (in clear view of Jones), intending to communicate thereby that he would like to go to see Flat-top Mountain. Jones, however, mistakes the drawing for one of a cowboy hat, and infers thereby that Smith would like to go to see Flat-top Mountain.

According to Buchanan, "what Smith intended to communicate, and all he intended to communicate, was *that he wants to go to see Flat-top Mountain*" (Buchanan,

p. 63). The author then goes on to claim that, unlike in Loar-case 1, here we seem to have no clear candidate for a MOP that the speaker had in mind and that the hearer failed to grasp, that is, "it is completely unclear what the MOP-involving proposition could be in this case" (ibid.) – but if that were true, then the MOP-based view would seem to have no resources for explaining why their communication appears to have been unsuccessful. The intentions-based view, on the other hand, easily explains what is going on in the story by invoking Smith's intention that Jones recognizes his drawing as a drawing of a mountain (and not of a hat).

As a first attempt of a reply, one could complain that Buchanan fails to consider the most obvious Fregean response: Jones thinks of the movie Flat-top Mountain *via* the MOP *the salient movie related to cowboy hats* instead of the MOP that Smith had in mind, *the salient movie related to a flat mountain*. That response could be based on the fact that pictures, just as much as linguistic expressions, can be associated with multiple semantic values, including reference and modes of presentation.[71] Thus, the response would continue, communication fails because Smith's drawing refers to Flat-top Mountain *via* the mountain-MOP but is taken by Jones to refer to that movie *via* the distinct cowboy hat-MOP. Jones would have gotten the right referent by means of the wrong MOP.

To be sure, this is not the only account of Buchanan's case available to a friend of MOPs. Indeed, even if, for whatever reason, one had suspicions about the idea of Smith's drawing being used to singularly refer to Flat-top Mountain,[72] a MOP-based theorist could suggest the following alternative account: Smith's drawing is a genre picture, a kind of depiction whose content is general (like a picture of a horse, but of no particular horse), and its content is the property of being a mountain.[73] It is by means of

---

[71] Hyman (2012) defends the application of the sense/reference distinction to pictures and other forms of depictions.

[72] One could, for example, rephrase Buchanan's story so that the intended communicated content were completely general, e.g. Smith could have intended to express his adoration for Western movies by means of drawing a mountain (commonly associated to the landscape of movies set in the wild west) whereas Jones thought that he was doing so by means of a drawing of a cowboy hat (equally associated to that type of movies). In that variation of the story, there would seem to be no singular content for the drawing to be a representation of.

[73] Analogously, one could say it represents the concept of a mountain. Genre pictures are unlike other paradigmatic instances of depiction, such as portraits, in that they refer to kinds of objects but not to particular instances of them. This means that a simple account of pictorial representation according to which they represent what they resemble would not be immediately applicable to them. For more about theories of depiction and genre pictures, see Hyman (2012).

figuring out that his drawing is a drawing of (the property of being) a mountain, that Smith hopes Jones will infer which movie he wants to go see. In Fregean terms, this means that the MOP of Smith's drawing is that of the property of being a mountain. One can, if one wants, say that this MOP either refers to the property of being a mountain or that it doesn't refer at all. Not much will hinge on this, since Jones' mistake can be accounted for by the fact that he fails to correctly recognize the drawing's MOP and, instead, takes its MOP as being that of the property of being a cowboy hat. If one believes that the correct MOP either referred to a property or that it didn't refer at all, then one gets the additional verdict that Jones not only failed to assign the correct MOP, but also the correct referent of the drawing. In any case, there seems to be no difficulty for a MOP-based view to account for why Jones has failed to understand Smith.

In summary: Jones misinterprets Smith's drawing as if it were of (the property of being) a cowboy hat, instead of (the property of being) a mountain. But misinterpretation just is, at least in a Fregean framework, the incorrect assignment of MOP to a representation. Thus, their communication fails because Jones fails to associate the correct MOP to Smith's drawing. Furthermore, given some additional assumptions about the drawing's MOP's referent, not only does Jones assign the wrong MOP to Smith's drawing, but also the wrong referent. The case can then be said to be analogous to one where a subject interprets an utterance of 'the [river]bank is muddy' as saying that some financial institution is dirty, and, *contra* Buchanan, not analogous to Loar-case 1 (where one at least gets the right referent).

That, I take, is a natural description of Buchanan-case which makes it clear that it is compatible with a MOP-based account of communication. Buchanan might have been assuming that MOPs somehow apply only to singular expressions, but that does not correspond to how the notion is employed in practice. Indeed, it seems that for any case that involves some type of misinterpretation – be it of an utterance, a drawing, or any other type of representational act – there will always be some easy way to account for them by means of a MOP-based view.[74]

[74] It seems that the same cannot be said for cases - such as those presented in Byrne & Thau (1996) in discussion with Heck (1995) – where there is no misinterpretation but where the correct interpretation is achieved by sheer luck or coincidence. It is less clear whether the MOP-based view possesses enough resources to deal with these cases or whether it would need to be supplemented with some additional epistemic constraints. I thank an anonymous referee for this commentary.

2.3 Interim conclusion: successful communication requires intention recognition

This is what we have so far: C1 fails as a criterion of successful communication because of cases such as Loar-case 1. In response to that, Loar suggested a MOP-based account of communication which, although coherent and able to account for all the relevant *data*, was shown by Buchanan to be excessively committal. Against Loar, Buchanan suggests C2, a principle which seems to be able to do all the work that Loar's MOP-based view was supposed to do and for a much lower theoretical price.

This could have been the end of the story if it were not for a recent paper by Unnsteinsson (2018) from where one can extract the following criticism: Buchanan's appeal to intention recognition is not adequately explanatory since it does nothing to explain why our intuitions on communicative success seem to change so drastically when we let go of one of the assumptions made in the previous Loar-cases – namely, that hearer and speaker are ignorant of some relevant identity fact. According to Unnsteinsson's proposal, one's account of successful communication must give a much more central role to the presence or absence of false distinctness beliefs such as those that Smith and Jones hold about the man on the TV and the man on the train.

3 The relevance of false distinctness beliefs: against Unnsteinsson (2018)

Notice what happens to one's intuitions about communicative success in Loar-cases as soon as we assume that the relevant subjects are enlightened with respect to the salient identity fact in the story:

(Loar-case 2): Suppose that it is mutual knowledge between Smith and Jones that the man being interviewed on television is someone they see on the train every morning.[75] Smith and Jones have not been engaging in any conversation when Smith abruptly says

[75] In an earlier version of this paper, it was merely assumed that Smith and Jones knew (and knew that each other knew) that the man being interviewed on television is a man they see every morning. However, as an anonymous referee aptly pointed out, this formulation of the story would be susceptible to familiar problems in case one of the subjects falsely believed that the other falsely believed that he falsely believed that the man on the television was distinct from the man on the train. The concept of *mutual knowledge*, taken from Schiffer (1972, p. 30), was specially devised to take care of similar types of cases. A and B mutually know that p if and only if (i) A and B know that p, (ii) A knows that B knows that p, (iii) B knows that A knows that p, (iv) A knows that B knows that A knows that p, (v) B knows that A knows that B knows that p, (vi) A knows that B knows that A knows that B knows that p, and so on *ad infinitum*.

‘That man is a stockbroker’, while pointing to the man on television; coincidentally, at that very moment Jones happened to be remembering the last encounter they had with that man on the train. Influenced by his own memory and failing to notice that Smith was pointing to the television, Jones takes Smith’s use of ‘That man’ as intending to bring about a memory of that man in the train. Now Jones, as it happens, has correctly identified Smith’s referent, since the man on television is the man on the train – but did not do so by means of the proper method.

Do we want to say that Loar-case 2 is a case of successful communication? Differently from Loar-case 1, where our intuitions weighed heavily in favor of communicative failure, Loar-case 2 is constructed in such a way that our intuitions, by themselves, do not seem to point *decisively* in either way. As Unnsteinsson (2018) suggests, a good way to begin answering this question is by reflecting on what Smith’s reaction would be if he noticed that Jones never realized he was pointing at the television. Here is one natural possibility about what would happen: nothing. Smith would be completely indifferent since his most general objective, e.g. letting Jones know some guy is a stockbroker, would be fulfilled regardless of the inferential process by means of which Jones reached that thought. To be fair, he could feel a bit annoyed and make it clear to Jones that he was actually pointing to the man on the television, but that would certainly be a tad pedantic, and a proper response from Jones would be ‘So what? It’s the same person!’.

Unnsteinsson takes these observations as evidence that the failure of communication in Loar-case 1 has got more to do with the fact that speaker and hearer hold a false belief about the distinctness the object referred by the speaker in the conversation (that the man on the train is distinct from the man on the television) than with the way by means of which the hearer interprets the speaker.[76] The author’s suggestion is that “it is in the nature of the speech act of singular reference that having specific false beliefs about identity can make it impossible for a speaker to perform the act properly”, such that “lacking such false beliefs at the time of the utterance can be a

[76] For the sake of simplicity, I will just assume that the subjects in our thought experiments do in fact hold the relevant false distinctness belief, but all of what I am going to say (as well as all of what Unnsteinsson says) is compatible with them merely suspending belief on the issue.

condition on the proper functioning of the underlying mechanism of singular communication" (p. 5).

One natural way to read Unnsteinsson's proposal is that lacking distinctness false beliefs must be taken as a necessary condition for communicative success.[77] A view of referential communication as a speech act with underlying normative constraints which preclude its proper functioning by confused speakers could have promising features. Nonetheless, I think it leads to clearly counterintuitive consequences.[78] For starters, one could insist that Loar-case 2 is an instance of communication failure, regardless of Smith's indifference. One could defend that view by claiming that sometimes our extra-communicative objectives get fulfilled even when, strictly speaking, our interlocutors have not grasped what we said. One could reinforce that view by arguing, as Evans (1982) and Heck (1995) once did, that the *raison d'être* of communication is the transmission of knowledge, something which is bound to be absent in every case where an element of luck is involved in the hearer's interpretation.

It is doubtful whether this knowledge-based view of communication would end up sounding convincing to everybody – as Pagin (2008, p. 30) critically says, "the claim that transfer of knowledge is 'the purpose of communication' strikes me as a piece of metaphysical speculation". In any case, there are even more pressing reasons why that criterion fails: it predicts communication failure every time speaker and hearer have a false identity belief, even when that belief is playing no immediate role in the

[77]Unnsteinsson is more closely concerned with providing necessary conditions for the success of the speech act of singular reference and does not discuss criteria for successful communication *per se*. My reading of that author is thus committed to a certain extension of his view under the plausible assumption that, if false distinctness beliefs would be disruptive of the success of the speech act of singular reference, then they would also be disruptive of the possibility of understanding them.

[78] Three things should be noted about my discussion of Unnsteinsson's view. Firstly, I am merely concerned with criticizing one particular thesis – that lacking false distinctness beliefs is necessary for communication – that is congenial, but, without further argument, not necessarily essential, to what this author defends in a brief discussion note. As is often the case with these short notes, Unnsteinsson might not have had enough space to fully develop his views and, in any case, it is not clear whether his other points are not compatible with my criticisms. Secondly, there is much of interest in his overall project – that of providing a speaker-based intentionalist theory of reference - that remains untouched by my discussion, such as his discussion of the optimal conditions of the speech act of singular reference and of the viability of an intentions-based view of reference in the face of the conflicting intentions confused speakers have. I refer the reader to Unnsteinsson's other works (2016, forthcoming) for more details. Finally, it might be that Unnsteinsson has a different methodological objective than I do. While I am concerned with an account of the folk notion of communication, he might be concerned with constructing a technical notion able to play a set of explanatory roles in his more general theory of reference. If that is the case, it is possible that he would try to resist arguments based on folk intuitions (see specially his forthcoming paper).

conversation. It strikes me as obviously true that we can sometimes successfully refer and communicate about people about whom we hold some irrelevant false distinctness belief: Lois Lane successfully talked about Superman almost every day even though she would laugh at the idea that his true identity was that of her boyfriend, Clark Kent.

In a couple of places, Unnsteinsson (p. 3, 5) suggests that the problem of false distinctness beliefs arises only when they are contextually salient or relevant. That sounds like a plausible enough idea because it is hard to deny that there are cases where subjects successfully communicate regardless of holding false distinctness beliefs that are contextually irrelevant, e.g. ancient Babylonians were able to successfully communicate about Venus in various contexts (during the day by means of 'Phosphorus', during the night by means of 'Hesperus') even though none of them knew that Hesperus is Phosphorus. In other words, that ancient Babylonians falsely believed that Hesperus and Phosphorus were distinct stars obviously did not interfere with the fact that they often had conversations about Venus.

What about Smith and Jones' belief that the man on the train is not the man on the television in Loar-case 1? One could argue that this was a contextually relevant belief since they were talking about the man on the train right before Smith made an utterance referring to the man on the television. So far, so good: Unnsteinsson's necessary condition is disrespected and communication does indeed fail. However, re-run Loar-case 1 with a slight modification: assume that Jones notices that Smith is pointing to the man on the television and thus that Jones doesn't make any interpretative mistake this time. I find it obvious that their ensuing communication would be successful regardless of their false and contextually salient distinctness belief – still, Unnsteinsson's criterion would output the opposite prediction. In summary, subjects can successfully communicate in the presence of false and contextually salient distinctness beliefs.

Maybe I'm not being completely fair to Unnsteinsson's notion of contextual relevance. Perhaps what Unnsteinsson means is that false distinctness beliefs only disrupt referential communication when the speaker intends to refer to one of the flanks of the belief and the hearer takes her to be referring to the other. In other words, perhaps Unnsteinsson is thinking of a case where (i) there are two ways W^1 and W^2 of singularly referring to the same individual, (ii) speaker and hearer falsely believe that W^1 and W^2 are ways of singularly referring to distinct individuals, (iii) speaker intends to refer to

that individual by means of W^1 and (iv) the hearer takes speaker to have referred to that individual by means of W^2. For illustration, W^1 could be *demonstratively referring to an individual as the man on the television* and W^2, *demonstratively referring to an individual as the man on the train*.[79] Plausibly, if all four of these conditions are satisfied, speaker and hearer will have failed to successfully communicate.

Even if it rings true, this idea will not be of much help to the view that lacking false beliefs is necessary for successful communication, since then the resulting view would simply boil down to the claim that false distinctness beliefs disrupt communication when the hearer fails to recognize the speaker's communicative intentions. That is just what is happening in the situation previously described: the hearer thought that the speaker intended to talk about the man on the train while actually she intended to talk about the man on the television. But that's already accounted for by means of C2. Why would we additionally need to give a central role to false distinctness beliefs in our criterion for successful communication if what really matters is whether the hearer gets the speaker's intentions?

In summary, Unnsteinsson's idea was supposed to help us understand why our intuitions become uncertain in cases where enlightened subjects misinterpret one another, such as in Loar-case 2. Against this author, I have argued that it is implausible to claim that subjects cannot successfully communicate about x when they hold false distinctness beliefs about x (even when these are contextually relevant), i.e. lacking distinctness false beliefs is not necessary for communication success. Furthermore, I have not found any adequate way of fleshing out what Unnsteinsson means by 'contextually salient' without making his criterion redundant.

Up until now, one could complain, I have only shown that lacking false distinctness beliefs is not necessary for communicative success. But that is compatible with a modification of Unnsteinsson's view that still gives a central role to the presence or absence of these beliefs. According to this hypothetical view, while it is not necessary that speaker and hearer be enlightened for them to successfully communicate, if they are indeed enlightened, then it is sufficient for the success of their communication that the hearer entertains a thought with the same truth-conditions as the speaker. In other words,

[79] Another example would be: W^1 is referring to Venus by means of 'Hesperus'; W^2, by means of 'Phosphorus'.

according to this view, successful communication is particularly easy for enlightened subjects, i.e. that C1 is true for pairs of speaker and hearer that hold no false distinctness belief of the matter at hand.

I will finish this section by arguing that C1 is false even for subjects that hold no false distinctness beliefs. This should be seen as providing more justification for accepting my claim that false distinctness beliefs should not have a central place in an account of successful communication. A quick example is enough to suggest why that is the case. Imagine, again, a case where it is common knowledge between Smith and Jones that the man on the train is the man on the television, but where Jones interprets Smith's utterance of

(3) That man [pointing to the man on the television] is that man [intending to refer to the man they saw on the train]

as if Smith had said

(4) That man [pointing to the man on the television] is that man [still pointing to the same man on the television]

I take it for granted that this would not be a case of successful communication. What this shows is that, even if speaker and hearer are enlightened, successful communication will not come so easily: C1 fails even for that restricted set of subjects. In conclusion, I have argued *contra* Unnsteinsson that one should not hold the absence of distinctness false beliefs to be a necessary condition for successful communication. I then have made the additional point that, even if one is only concerned with subjects who hold no false distinctness beliefs about the subject matter of their conversation, it is not the case that entertaining a thought with the same truth-conditions as the speaker's is a sufficient condition for communicative success. My overall aim was arguing for the thesis that false distinctness beliefs should not play a central role in one's account of successful communication.

Regardless of all that, we still haven't explained one of the main intuitions behind Unnsteinsson's suggestion, that is, we still need to provide some kind of explanation for

why is it that our intuitions change so clearly as soon as we re-describe Loar-case 1 as Loar-case 2, i.e. when ignorance gives way to enlightenment. If my objective of rescuing C2 is to be achieved, there is still some work to be done.

4 Enlightenment, linking and merging

As I have remarked, one could accept all of the previous considerations but still wonder why is it that, when we enlighten the hearer and speaker in a Loar-case, as we did in Loar-case 2, our intuitions suddenly become much more sympathetic towards the verdict that they have successfully communicated regardless of the hearer failing to properly grasp the speaker's communicative intentions. That would mean C2 is subject to clear counter-examples.

Loar-case 2 might be less than ideal as a case study because it involves an explicit element of luck in the way the hearer reaches his interpretation. Regardless of what one thinks about communication and luck, it will be useful to analyze a variation of the same case which does not include an accidental strike of luck.[80] It will then be harder to deny that we can have communicative success even when a hearer seems to have failed to grasp the speaker's communicative intentions:

(Loar-case 3) Suppose that it is mutual knowledge between Smith and Jones that the man being interviewed on television is someone they see on the train every morning and about whom, in that latter role, they have just been talking. Indeed, they have been talking about that man for hours, constantly alternating between referring to him by pointing to his image on the television or by remembering their encounters in the train. Smith says 'That man is a stockbroker', while pointing to the man on television; unaware of Smith's pointing gesture, Jones takes Smith to have been invoking a memory of one their encounters in the train (perhaps of one day in which the man was dressed as a stockbroker). Now Jones, as it happens, has correctly identified Smith's referent, since

[80] I have been purposefully avoiding any discussion of the role of luck in Loar-cases and its relevance for criteria of successful communication. As far as the recent literature on the topic goes, there just is no consensus on the question about whether successful communication can be lucky. One tradition that includes Evans (1982), Heck (1995) and Peet (2017) says it cannot. Another tradition including Byrne & Thau (1996), Paul (1999), and Pagin (2008) says it can. At the end of the day, I do not think it is necessary, for the objectives of this paper, to try to settle on an answer to this question here.

the man on television is the man on the train – and was not merely lucky to have done so.

It is hard to deny that Jones has understood what Smith said and that their communication was successful. Whether it matters or not, Jones was not merely lucky to identify the referent of Smith's utterance. On the contrary, it is obvious that he knew who Smith was talking about, regardless of failing to notice his pointing at the television. Regardless of all that, one could argue that C2 would *prima facie* entail that their communication was not successful since the hearer seems to fail to grasp the speaker's communicative intention. In the final section of this paper, I will provide a brief explanation of how one could account for Loar-case 3 without departing from C2.

There is an easy way to explain the communicative success in scenarios such as Loar-case 3 where the subjects are enlightened but still *seem to* misinterpret each other. In order to do so, I suggest we borrow the notions of linking and merging from Strawson (1974, p. 51-56) and their further developments by Recanati (2012, p. 42-53). The idea, in a nutshell, is that identity judgments be understood as enabling the flow of information between representations that the subjects previously thought referred to distinct people. For the purposes of our paper, linking and/or merging representations can be conceived as processes that have an effect on a subject's communicative intentions with these representations. When a speaker acquires an identity belief about what she previously thought were two distinct things, the conditions of fulfillment of her communicative intentions about that thing will also expand so as to cohere with her new state of mind. Let us see how that would work in practice.

Assume that there is a time *t* previous to the interaction described in Loar-case 3 when Smith and Jones are still unsure about whether the man on the train is the man on the television. At *t*, Smith and Jones take the world to be such that there are two distinct individuals satisfying those predicates and strive to isolate the information they acquire about each of them. At that point of the story, when Smith makes an utterance about e.g. the man on the train, the fulfillment of his communicative intention depends on Jones taking him to be remembering the man on the train (and not, for example, pointing at the man on the television). When these subjects learn that the man on the train is the man on the television, they are rationally required to update their mental states by means of the

mechanisms of linking and merging. In other words, this means they no longer have a reason for keeping the information about the man on the train insulated from information about the man on the television. In response to that, they should either create a bridge between their information repositories or merge them, so that all of the information can be stored in one and the same place.[81] From that point onward, Smith's communicative intentions become fulfillable by his interlocutors taking him to be intending to refer to the man on the train. Thus, his communicative intentions expand so as to encompass distinct ways of referring to what he now knows is one and the same person.

Now we can see why false distinctness beliefs matter to our intuitions of communicative success. When we acquire information about one and the same thing but are unaware of that fact, we are rationally required to keep that information in distinct 'compartments' until we make sure that they come, indeed, from the same source. When we discover that we are receiving information from the same thing from different non-coinciding means, we then have the option of merging the information into one and the same compartment or linking the information contained in distinct compartments in order to allow for their free flow (by expanding the fulfillment conditions of our communicative intentions). To be fair, Unnsteinsson could try to incorporate these linking and merging mechanisms to his account – but if he did, then he would owe us an explanation for why false distinctness beliefs are fundamentally, and not only derivatively, explanatory.

Thus, when Smith and Jones become aware of the relevant identity fact, they link or merge the information about the man on the train and the man on the television. Most relevantly, this means that, contrary to what we thought, C2 is actually compatible with ascribing communicative success to Loar-case 3. When Jones interprets Smith as intending to refer to the man on the train, one could argue that he thinks a thought with the same truth-conditions as Smith and that he has also recognized Smith's communicative intentions, since, after undergoing the mechanisms of linking and merging, thinking of the relevant man as *the man on the train* or as *the man on the television* are equally correct ways as far as Smith's communicative intentions are

[81] I do not intend that my folder, repository or compartment-talk to be taken as anything more than metaphors for how we organize the information we possess.

concerned. I expect those brief considerations to pave the way for an explanation of Unnsteinsson's *data* without letting go of C2.

5 Final remarks

We started with the simplest criterion for successful communication (C1) and argued that it does not output the correct predictions for Loar-cases. I then assessed Loar's MOP-based account of the problem in light of Buchanan's critical remarks. My conclusion was sympathetic to Buchanan's with the exception of my point that, even if an intentions-based account of communicative success (C2) is, as far as this limited set of *datum* is concerned, less theoretically committal, a MOP-based account is still compatible with everything we've seen so far.

In the following section of the paper I then went on to argue against Unnsteinsson's recent suggestions to the effect that the failure of communication in Loar-cases is to be explained by the presence of contextually relevant false distinctness beliefs about the object being referred to by the speaker. In order to achieve that conclusion, I argued that there can be communicative success even if speaker and hearer hold contextually relevant false distinctness beliefs. I then tried to develop Unnsteinsson's suggestion in a different way but argued that it was bound to collapse into the claim that false distinctness beliefs only disrupt communication when the hearer fails to grasp the speaker's intention – and that this would make his suggestion theoretically redundant. I then finished this section by motivating the claim that, even if speaker and hearer are enlightened, it is not sufficient for successful communication that the hearer entertains a thought with the same truth-conditions as the speaker.

Finally, I turned my attention to Unnsteinsson's remaining challenge: how would a view like C2 account for the fact that we are disposed to ascribe communicative success to cases, like Loar-case 3, where the hearer seems to fail to grasp the speaker's communicative intentions? My reply, admittedly tentative, employed the notions of linking and merging in order to argue that, when thinkers discover that they were acquiring information from a single object by means of two distinct sources, they tend to link or merge this information, and that something like this could very well explain what is going on in the cases Unnsteinsson calls attention to.

Naturally, there are plenty of open questions we did not have time to touch upon. One of these is: can successful communication be lucky? Just as a hearer might luckily manage to get the truth-conditions of the speaker's thought right, she might also luckily manage to recognize the speaker's communicative intentions. That would mean that C2 is – as much as C1 – subject to cases where a hearer complies with it by means of sheer luck. Then, answering whether C2 is an adequate criterion of successful communication at all takes us back to the question of whether communicative success can go hand in hand with luck – a question I have not even pretended to be able to answer.

A distinct challenge to C2 draws upon the fact that, as is common for Gricean-inspired views, the resulting account might end up being psychologically too demanding. It is a well-known fact that thinkers usually have different perspectives on the subject matters they think about. Likewise, it is often vague which inferential route to the referent is intended by particular referential speech acts (Peet, 2016). Requiring that the hearer thinks of the referent in the exact same way as the speaker, or that the hearer magically discover the exact inferential route hidden behind the speaker's utterance might quickly lead one to a criterion which is rarely, if ever, satisfied in the real world.

Regardless of the amount of work ahead of us, I expect to have at least made a good case to the effect that progress on the issue of successful referential communication should not give fundamental importance to the presence or absence of false distinctness beliefs, and that it should drive off the idea that, even for some restricted cases, identity of referential content is enough. Communicating well, just as understanding well and holding the same belief, seems to be a hyper-intensional notion that requires more than an actual match of referents. How exactly we should develop that thought, however, remains a question for future investigation.

References

Buchanan, R. (2014) Reference, Understanding, and Communication, *Australasian Journal of Philosophy* 92/1: 55–70.

Byrne, A. and Thau, M. (1996) "In Defense of the Hybrid View." *Mind* 105 (1996): 139-49

Evans, G. (1982) *The Varieties of Reference*, ed. John McDowell. Oxford: Clarendon Press.

García-Carpintero, M. (2016) Token-Reflexive Presuppositions and the De Se. In *About Oneself: De Se Thought and Communication*, edited by Stephan Torre and Manuel Garcia-Carpintero, 179–199. Oxford: Oxford University Press.

Heck, R. (1995) The Sense of Communication. *Mind* 104 (1995): 79-106.

Hyman, J. (2012) Depiction, *Royal Institute of Philosophy Supplement* 71:129-150.

Kaplan, D. (1990) Words, *Aristotelian Society Supplementary* Volume 64 (1):93-119.

Loar, B. (1976) The Semantics of Singular Terms, *Philosophical Studies*, 30/6: 353–77.

Paul, M. (1999) Success in Referential Communication. Dordrecht: Kluwer Academic Publishers.

Pagin, P. (forthcoming) 'When does communication succeed? The case of general terms' in Teresa Marques and Åsa Wikforss (eds.), *Shifting Concepts*, OUP, (expected 2019).

Pagin, P. (2008) What is communicative success? *Canadian Journal of Philosophy* 38 (1): pp. 85-115.

Peet, A. (2016) Referential Intentions and Communicative Luck, *Australasian Journal of Philosophy* 95/2: 379–84.

Recanati, F. (2012) *Mental Files*. Oxford University Press.

Recanati, F. (2016) Indexical Thought: The Communication Problem. In *About Oneself: De Se Thought and Communication*, edited by Stephan Torre and Manuel Garcia Carpintero, 141-178. Oxford: Oxford University Press.

Salmon, N. (1986) *Frege's Puzzle*, Cambridge, MA: MIT Press.

Schiffer, S. (1972) *Meaning*, Oxford University Press.

Strawson, P. (1974) *Subject and Predicate in Logic and Grammar,* London: Methuen.

Unnsteinsson, E. (2016) Confusion is Corruptive Belief in False Identity. *Canadian Journal of Philosophy* 46 (2):204-227.

Unnsteinsson, E. (2018) Referential Intentions: A Response to Buchanan and Peet. *Australasian Journal of Philosophy* 96 (3):610-615.

Unnsteinsson, E. (forthcoming) The Edenic Theory of Reference. *Inquiry: An Interdisciplinary Journal of Philosophy*:1-33.

5 Relationism and the Problem of Publicity

(with Victor Verdejo[82])

Abstract

According to a recently developed family of relational views, whether two concepts C_1 and C_2 are the same or shared is a matter of an external relation in which (tokens of) C_1 and C_2 stand. In this paper, we (i) highlight the chief contributions of Relationism in the elucidation of concept sameness, (ii) present a set of arguments to the effect that relational accounts of concept sameness fail to accommodate a substantive notion of concept publicity, and (iii) offer a diagnosis of this result: whereas, as has been pointed out by relationists, non-relational approaches to concepts often fall short of concept shareability, Relationism puts forward a form of concept shareability that itself falls short of concept publicity. We conclude that the strengths of non-relational approaches will also need to be considered in order to fully capture what it means for a concept to be public.

1 Introduction

A number of groundbreaking works have recently advanced the view that concept sameness must be understood in relational terms. According to this view, whether two concepts C_1 and C_2 are the same or shared is a matter of an external relation in which (tokens of) C_1 and C_2 stand. The class of approaches offering variations of this view goes under the head of 'Relationism'.[83] While we do not deny the import of the proposed relations as fundamental aspects of cognition and communication, in this paper we target one of the main tenets of this momentous position. In particular, we set out to show that relational accounts of concept sameness fail to accommodate a substantive notion of concept publicity. We take this result to be revealing in ways that are challenging for the prospects of the relationist project as a whole but especially so in light of the fact that capturing concept publicity is, as we shall see, one of its principal motivations.

In order to fulfil this task, we first highlight the chief contributions of Relationism in the elucidation of concept sameness (Section 1), and go on to provide a battery of

[82] Víctor is a professor of philosophy at the Universitat de València, Spain.

[83] The first monograph-length defence of the view is Fine's *Semantic Relationism* (2007), but its roots trace back at least to Taschek (1995). See Gray (2017) and below for detailed discussion.

arguments to the effect that Relationism fails to deliver sufficient and necessary conditions for concept publicity (Section 2). We will conclude that the relational account of concept sameness does not yield sameness for public concepts. In a more speculative fashion, we also propose a diagnosis of this result and its significance (Section 3). Whereas non-relational approaches to concepts, as standardly conceived, fall short of concept publicity because they fail to guarantee concept shareability, relational approaches also fail in achieving that objective, albeit for different reasons. More specifically, Relationism's criteria for concept sameness oscillate between being too strict for concept publicity – disallowing unrelated thinkers from sharing the same thoughts – and too loose – identifying the concepts of thinkers that, for independent reasons, we would like to distinguish. In other words, Relationism puts forward a form of concept shareability that itself falls short of concept publicity. We conclude that, while neither type of approach can rightly be said to have fully captured concept publicity thus far, the strengths of non-relational approaches will also need to be considered in order to do so.

2 Relationism and Publicity

Concepts are the atomic constituents of thoughts. As such, they are usually expected to play a number of theoretical roles, some of which are philosophically contentious, some of which are too basic to be contested. One central example of a basic role concepts are supposed to play is accounting for the cognitive significance of propositional attitudes. This we may dub the cognitive-significance role or CS-Role for short. Because concepts fulfil this role, they can feature prominently in explanations of why some beliefs can be cognitively distinct regardless of having the same truth-conditions, e.g. the belief that Istanbul is Constantinople and the belief that Istanbul is identical to itself. The natural concept-based explanation of that cognitive difference is that the first belief is constituted by a thought that contains two concepts, the second, only one twice over.

Another basic theoretical role of concepts – one that is central to this paper – is that concepts are, as it were, the glue connecting the mental life of thinkers. As emphasized by Fodor (1998, 28), "[c]oncepts are *public*; they're the sort of things that lots of people can, and do, *share*" (emphasis in original). This we may call the publicity

role or P-Role as an abbreviation. As hinted in Fodor's dictum, publicity is clearly more than mere shareability. The publicity of concepts does not merely require that they be, in principle, shareable, but that they be indeed generally shared. We may cash this idea out in terms of a desideratum for theories of concepts: any theory according to which concepts are not merely, in principle, shareable, but also somehow guaranteed to be shared in a community of thinkers scores more points.

Relationists generally believe that one strong reason to accept their views is that, not only are they able to account for the cognitive significance of propositional attitudes (their CS-Role, that is), they are also better equipped to account for concept publicity and sharing than the others (the P-Role in our terminology).[84] But to see exactly why we need to characterize Relationism more accurately. Various types of relational theories have been proposed in the last couple of decades. What they all have in common is not always easy to discern but can initially be summarized as the idea that whether two concepts C_1 and C_2 are the same does not supervene on C_1 and C_2's intrinsic properties, but rather on whether they, or more precisely their tokens, are externally related. A relation is external, in the relevant sense, iff it is not reducible to a relation between the properties intrinsically possessed by each concept.

We do not even pretend to have a fully fleshed-out view of intrinsicality, something most desirable when invoking the internal-external relation distinction. However, a clear analogy might be all we need in this context. Thus, *being taller than* is a paradigmatic example of an internal relation. It is wholly determined by the intrinsic properties of its relata. Indeed, one could say that the holding of this relation between two individuals amounts to nothing over and above each of them having a certain height. The relation of *being south of,* by contrast, is a paradigmatically external one. In order to assess whether two objects are thus related, it is not enough that we have access to the individual's intrinsic properties – we need to check the world around them. It does not matter that, by exploiting logic's expressive power, one can turn any relational property into a monadic property, e.g. if *x* is to the south of *y*, then *x* possesses the monadic property *being south of y*. Transforming relations into monadic properties in this way does not alter their internality or externality. If the transformed relation was external,

[84] A vivid illustration of this we find in the titles of Onofri's (2018) "The Publicity of Thought" and Prosser's (2019) "Shared Modes of Presentation".

then the corresponding monadic property will not be based on the individual's intrinsic properties.

We take the above example to spell out the meaning of the proposed relationist slogan "concept sameness is an external relation".[85] By contrast, non-relational views are those according to which two concepts C_1 and C_2 are the same whenever C_1 is intrinsically associated with some property P_1, C_2, with some property P_2, and P_1 is identical to P_2. For example, P_1 and P_2 could be the concepts' syntactic type, associated definite description, way of thinking, perspective or understanding conditions. The key characteristic of non-relational views is that conceptual sameness boils down to the properties that concepts individually possess or is, to use Prosser's apt phrase, "a coincidence of individual achievements" (Prosser 2019, 3). Relational views, on the other hand, might very well agree that a concept possesses syntactic and semantic properties, but will argue that whether two token concepts are the same is not determined by them. Thus, we regard Relationism as follows:

> **(Relationism)** Two concepts C1 and C2 are the same iff their tokens are related by R, where R is an external relation.

Relationism concerns therefore concept sameness as predicated of concept types in terms of a relational condition between concept tokens. Every particular relational theory, however, will have something different to say about how to understand R. For ease of exposition, we may compress the claim that any such relational condition between tokens of C1 and C2 is fulfilled in the claim that C1 and C2 are 'relationally connected'. Thus, according to any relational theory, two concepts are the same iff relationally connected.

Now, to return to our previous question, why would one believe that Relationism is better positioned to capture publicity than the traditional, non-relational view? The basic answer is that traditional accounts of what it is to share a concept seem too demanding. They suggest that people would rarely, if ever, share their concepts. If, for

[85] It is thus no accident that Fine (2007) uses the term 'Intrinsicalism' to denote the family of views opposed to Relationism. Gray (2017) also defines Relationism as the view that sameness of representation does not supervene on *intrinsic representational features*.

example, the way of thinking or perspective by means of which a thinker perceives or regards an object is an intrinsic property constitutive of a concept C, then, on the traditional account, another thinker will only have the same concept C if she also exploits the very same way of thinking or perspective. However, ways of thinking or perspectives are, plausibly, highly variable from subject to subject and even continuously for the same subject across time. It follows that two thinkers would rarely if ever share the concepts constituted by ways of thinking or perspectives on the traditional view. In terms of the desideratum that we had earlier presented, the theory just sketched guarantees that concepts are, in principle, shareable but not that they are actually shared in communities of thinkers or across communities.

In response to this shortcoming, Relationism offers interesting new routes to explaining how thinkers with very different ways of thinking or perspectives about a certain subject matter nonetheless share their concepts in communication and rational interaction. The key idea is that concepts used by different thinkers can be relationally connected – and hence be the same on relationist standards – regardless of any differences in way of thinking or perspective.

There are a number of ways in which the relationist story about shareability can be fleshed out. For instance, some think that relational connectedness should be modeled as a primitive semantic-coordination relation that determines, beyond any intrinsic semantic properties, which elements of content in a sequence of representations are represented as the same (Fine 2007, Pinillos 2011). Others would propose that a simpler view could be reached by taking a formal or syntactic kind of relation as primitive (Cumming 2013, Heck 2012, 2014). According to Cumming's (2013) formal approach, for instance, concept sameness is a matter of the holding of some conventional coordination relation between thinkers' co-referential representations in a one-to-one fashion. Consider two subjects, *a* and *b*, and call A_1 the concept *a* uses to refer to O, and B_1 the one *b* uses for the same purpose. Is *a*'s concept (or concept type) the same as *b*'s? According to Cumming, this depends on which conventions *a* and *b* put into place. For instance, if *a* commits to tokening A_1 in response to *b*'s tokens of B_1 and vice-versa, then $A_1 = B_1$. As expected, however, the holding or not of that convention is not determined by any of A_1 and B_1's intrinsic properties. In a different context, the chosen convention could be distinct.

Although our arguments below are supposed to apply to any version of Relationism, here we will often explicitly target a conception that contrasts with the semantic and formal kinds just surveyed and already singled out in the literature (see especially Gray 2017). This conception, which has been quickly gaining in popularity, we may dub 'epistemic relationism'. According to it, relational connectedness is, first and foremost, an epistemic relation between thinkers and token concepts or representations, and only derivatively a semantic or formal one. One clear representative of this view is Onofri (2018). According to this author, two concepts are the same when their owners know that their tokens co-refer. This view too seems able to account for concept sharing without satisfying any of the presumably demanding criteria proposed by the traditional non-relational views.

Many epistemic relationists in our sense take relational connectedness to consist of a more subtle form of epistemic relation, namely, *trading on identity* (Dickie and Rattan (2010), Schroeter (2012), Prosser (2019)). Thus, one trades on the identity of two concepts or thoughts when one treats them as representing something as the same without needing to establish it by means of an intermediary identity premise. Campbell (1987) convincingly argued that, unless some inferences involve trading on identity in this sense, we quickly fall into a very pernicious regress. The argument is simple: if the validity of every inference required identity premises connecting the referents of the inference premises with each other, then, since those identities would themselves become premises of the original inference, we would need additional identity premises connecting the referents of those identities with those of the original premises, and so on *ad infinitum*. Building on these considerations, some relationists go on to suggest that concept sameness is the result of trading on the identity of concepts' reference – viz. in communicative exchanges and in belief retention – without requiring the identity of hardly shareable intrinsic properties of concepts.

We have now seen several ways in which relational views seem to be well-equipped to deal with the publicity of concepts. In particular, and unlike traditional non-relational approaches, they seem to satisfactorily accommodate concept sharing in spite of differences in variable intrinsic properties exploited by different thinkers. However, a substantive notion of concept publicity is plausibly much more than concept sharing

beyond individual differences. There are, as we now turn to see, other aspects of publicity that also need to be handled but escape the resources of Relationism.

3 Relationism and Publicity, Unabridged

There are two kinds of challenges that would seem to threaten the relational project when it comes to a substantive notion of concept publicity. Firstly, the conditions that Relationism delivers for concept sharing do not actually suffice for the sharing of public concepts. They are conditions that, if satisfied, signal a significant connection (semantic, syntactic or epistemic) between concepts that, nonetheless, publicity requires to be distinct. Secondly, some central ordinary cases in which the sharing of public concepts is available are cases where a relational condition is decidedly absent. These cases indicate that relational conditions are not necessary for the publicity of concepts. Let us take each kind of challenge in turn.

3.1 Relational conditions are not sufficient for publicity

Relational conditions are, we suggest, not sufficient conditions for the publicity of concepts and thought. Here we would like to illustrate the point through reflection on three plausible requirements of a substantive notion of concept publicity: type classification, generalizable intentional explanation and transitivity.

Type classification. One seemingly uncontroversial ingredient elicited by the idea that concepts are public is that they can be sorted into general types. On this account, to say that concepts are public or publicly accessible is to say that people have the capacity to entertain, not merely the same concepts, but the same general types of concept. There are probably various ways of characterizing the relevant notion of concept type at issue. To a first approximation, we may take the types in question to be carved up in alignment with semantic types of expression.

For instance, at least since the work of Kripke, Kaplan and other representatives of the so-called ‘New Theory of Reference’, it is customary to suppose that some expressions – e.g. indexicals and, arguably, bare and complex demonstratives and proper names – are directly referential whereas others – e.g. definite descriptions and quantifier phrases – pick out objects only indirectly. The distinction separates out two types of

expression semantically. The directly referential type includes terms that stand in a close or immediate relation to the objects they refer to, whereas the non-directly referential or descriptive type comprises expressions involving reference through some sort of medium, such as a reference-fixing description. If correct, this divide as standardly formulated suggests that concepts expressed by means of terms differing in semantic type belong to different general conceptual types and are indeed themselves distinct.

It is easy to see how this picture spells trouble for relational approaches. For sameness of concept must surely entail sameness of general conceptual type. However, the kind of conceptual sameness tendered by relational accounts may hold quite independently of semantic types of expression. Relationism opens thus the door to cases in which two (co-referring) concepts expressed with different semantic types (viz. an indexical and a definite description) are deemed of the same conceptual type insofar as relationally connected. On the assumption that semantic types of expression correspond to general concept types, this result is problematic.

Consider, to illustrate, the proper name "Hesperus" and the definite description "the heavenly body (actually) visible in the evening". In appropriate contexts, the relationist contends, these terms express the same concept. This would be the case, for instance, if subjects know that these terms both designate Venus (Onofri), or exhibit some form of trading on the identity of their reference (Dickie and Rattan, Schroeter, Prosser), or if the terms in the context are formally or semantically coordinated (Cumming, Fine, Heck). The important point here is that, according to Relationism, in the contexts in which "Hesperus" and "the heavenly body (actually) visible in the evening" are relationally connected, the terms must express the same concept. But this result is counterintuitive in light of mainstream approaches to meaning and thought that discriminate between directly referential and non-directly referential types. It strongly suggests that a relevant relational condition may be satisfied in regard to two concepts when publicity requires them to be, not merely distinct, but indeed of distinct general types.

Some relationists may counter that it is never the case that a proper name is relationally connected to a definite description. Take the view that relational connectedness consists of some form of trading on identity. Some could argue, for example, that the following inference scheme has no valid instances:

1. Hesperus is F
2. The heavenly body (actually) visible in the evening is G

Conclusion. Something is both F and G

If correct, this would indicate that there cannot be trading on identity between a proper name and a definite description, and that such inference schemes are never valid unless incremented with an identity premise (of the form "N = the actual M" for any names N and definite descriptions M). More generally, some could claim that there is never trading on identity between concepts of distinct semantic types.

The problem with this line of response would then be that of explaining, non-question-beggingly, why inference schemes like the one above cannot have valid instances. This cannot simply be taken as a basic fact. One also cannot say that these inferences are not valid because the singular concepts in each premise are distinct, since this is precisely the fact that trading on identity should explain. Indeed, the notion of trading on identity is a highly technical one and its application conditions are a matter of philosophical dispute. Even if Campbell (1987) showed that some inferences must involve trading on identity, he did not put forward a clear cut criterion for identifying which. Therefore, unless the relationist authors who give a central role to this or other inference-constraining notions can explain why that inference is invariably dependent on a hidden identity premise, they will be exposed to the counterintuitive consequence just pointed out.

An objector may also worry that the problem thus far presented concerns a particular classification into general types which might itself be challenged. However, one need not resort to any particular type distinction in order to put the point across. It suffices to accept that independent research in semantics, linguistics and psychology do make it plausible, and certainly conceivable, that concepts come in general types.[86] One's account of shareable concepts is thus prima facie undermined if insensitive to, or even

[86] Emar Maier for instance proposes to abandon the semantic distinction between reference and description (in philosophy) and the distinction between pronouns and R-expressions (in linguistics) in favor of the more encompassing distinction between definites and indefinites (Maier 2015). The argument in the main text replicates, *mutatis mutandis*, for these and alternative classifications.

incompatible with, such general type distinctions. But that is precisely what seems to happen if Relationism is true.[87]

Intentional explanation. Type classification is hardly the only difficulty that relationists face in cashing out a substantive notion of publicity. Let us turn now to the characteristic kind of explanation in which concepts and other mental representations feature, namely, intentional explanation of cognition and behavior. As Dilip Ninan has noted, it is reasonable to expect that such explanations be generalizable:

> A good explanation ought to be something that generalizes, something that leads to by-and-large correct predictions in new cases. If we know that Sally went to the zoo because of her beliefs and desires, then if we learn that Sam has those same beliefs and desires, then it is reasonable for us to expect that Sam too will go to the zoo, so long as other things are equal. (Ninan 2016, 101)

We take Ninan's point in this passage to capture a widely held understanding of the way intentional explanation is supposed to work. On this picture, publicity requires concepts to be the sort of thing that is associated with cognitive or behavioral patterns, where such patterns are generalizable both synchronically – across subjects – and diachronically – for the same subject across different contexts or times.

It is questionable that relational views can accommodate this aspect of publicity. A straightforward corollary of the above is that if C_1 and C_2 are the same concept, then they must play exactly the same explanatory role in the intentional explanations where they feature, i.e. if two subjects have the same beliefs and desires, then one can infer that they will perform the same actions *caeteris paribus*. Relationism is at odds with this. The reason is that relational connectedness does not necessarily march with sameness of explanatory role. This is problematic. If relational sameness does not guarantee sameness of explanatory role, it will lead to systematically unwarranted generalizations of

[87] Recent publications examine whether 'concept' has a univocal meaning in different disciplines, but the question is still highly polemical (e.g. Machery 2009, Löhr 2018). If Relationism takes a stand on it, regardless of being a theory devised to answer to other concerns, then that might very well be taken as an unfortunate conclusion.

explanations that invoke the same concepts. Alas, concept publicity in intentional explanations does not get off the ground if the same concepts do not secure sameness of explanatory contribution.

One vivid illustration of this hurdle is given by the oft-cited phenomenon of the essential indexical (Perry 1979). Barring sceptic drifts (e.g. Cappelen and Dever 2013), there is a wide consensus that concepts (or beliefs) involved in the use of indexical terms – such as 'I', 'now' or 'here' – are indispensable or explanatorily primitive in accounting for the way in which one may think and react to a state of affairs. For instance, in order to explain one's running to a meeting starting at noon, we may invoke the thought one would express by 'now is almost noon!' but not merely the thought one would express by '11:45 am is almost noon!' even if 'now' refers to 11:45 am. Thus, any complete explanation of an agent's behavior must involve what Perry called 'self-locating' beliefs or concepts (cf. Perry 1979, esp. 4-5; Perry 2006).

It is unclear whether Relationism can live up to the existence of self-locating concepts or, for that matter, any class of concepts involving distinctive or irreducible explanatory features. Take again the self-locating concept typically expressed by 'now'. The problem is plain once we realize that relationists are forced to admit that, in at least some cases, the concept expressed with 'now' to refer to t would be just the same as the concept expressed by any other t-referring term insofar as a semantic, formal or epistemic relation is satisfied. Consider for instance the view that knowledge of co-reference or some form of trading on identity is what constitutes concept sharing in a context. There will be indefinitely many cases in which we would seem to know the co-reference or trade on the identity of the reference of the concepts expressed with 'now' and other non-indexical terms such as 'noon' or '12.00 pm'.[88] But if the Perrian insights are at all valuable, it would seem that in such contexts too there is a persisting – indeed essential – difference between the concepts expressed by 'now', on the one hand, and 'noon' or '12.00 pm', on the other. No explanation of a thinker's running-off-to-a-meeting

[88] The potential counter-argument that we never trade on the identity of an indexical and a non-indexical term can be responded by the discussion in the previous section.

behavior would be complete without mentioning some belief containing the concept that she would express by 'now'.[89]

At this point, a sympathizer of Relationism might complain: to claim that two token concepts expressed with an indexical and a distinct co-referring non-indexical term are explanatorily equivalent in a context is not yet to disavow the significance and explanatorily distinctive character of self-concepts. On a relational account, the concepts expressed by 'now' and 'noon' only have different behavioral import when they are not relationally connected, e.g. when a thinker does not know that they co-refer. This would explain why Perry-cases typically involve a thinker who is unaware that e.g. the current time is noon. On the other hand, the relationist could continue, in scenarios in which these concepts are relationally connected, they have just the same behavioral import and are intersubstitutable in an intentional explanation.

However, this reply overlooks the fact that a complete explanation of the behavior of someone for which the concepts expressed with 'now' and 'noon' are relationally connected would *still* have to include a 'now'-concept, namely, the belief that she would then express by 'now is noon', without which it would not make sense, say, to run to the meeting after overhearing 'the meeting starts at noon'. The same is not true of a 'noon'-concept. This shows that the concepts expressed by 'now' and 'noon' are not, after all, intersubstitutable in the intentional explanations. More generally, the relationist lacks the resources to explain why behavior explanations seem to necessarily require the involvement of self-locating beliefs or concepts. This is just as expected, for to fully acknowledge this requirement would be to acknowledge the existence of concepts – viz. self-concepts – whose explanatory features can be described as intrinsic features and hence irrespective of the relations in which they enter towards other concepts in a context.[90]

---

[89] Prosser (2015, 2019, §6) proposes an elaborated account in which differences in what he terms 'manifested belief' (about the subject's being in a certain subject-environment relation) – as opposed to the 'stated belief' (roughly, the referentially specified belief) – is what explains the difference in behavior associated with indexicals such as 'here'/'there' or 'now'/'then'. The problem we are discussing reappears, however, with respect to the elucidation of the notion of 'manifested belief'. If we follow Prosser's suggestion, the relevant publicity presumption will be that sameness of concept or mode of presentation yields sameness of manifested belief. But this is precisely what is in question in the cases envisaged in the main text.

[90] Richard Heck seems to acquiesce to this line of objection in connection with demonstrative thought and, especially, indexicals (2012, 159-162) but takes this to be, at any rate, a 'small victory' for the non-

Transitivity. Most authors – including a number of relationists (e.g. Cumming 2013, Schroeter 2012, Onofri 2018) – are ready to accept that transitivity is a capital constraint for any satisfactory account of concept sameness and individuation. As Schroeter writes, using the label '*de jure* sameness' for sameness of concept,

> ordinary reasoning seems to commit us to the transitivity of *de jure* sameness in thought and talk. For instance, when you rely on standing beliefs about tigers in a stretch of conscious reasoning, you're not just committed to those beliefs pertaining *de jure* to the same topic. You're also implicitly committed to those beliefs pertaining *de jure* to the same topic as the past judgments from which they derive, and to the other past attitudes on which those past judgments were based – even if you no longer remember those attitudes. (Schroeter 2012, 17, ff. 18)

More than implied by our ordinary reasoning, however, the transitivity of concept sameness seems to open the way for a substantive vindication of concept publicity. A non-transitive notion of sameness would seem to jeopardize the very idea of concept publicity or, to put it in Fodor's terms, the idea that all sorts of concepts "are ones that all sorts of people, under all sorts of circumstances, have had and continue to have" (Fodor 1998, 29). The reason is surely obvious: if the fact that C_1 is the same as C_2 and C_2 the same as C_3 does not *ipso facto* make it true that C_1 is the same as C_3, then there might be a situation in which, even if I might share all my concepts with someone in the community, they quite generally fail to be the same concepts as those of my fellows. As previously observed, a substantive vindication of publicity does not merely require that it be possible that subjects share their concepts. It requires, more precisely, that indefinitely large numbers of subjects do share the very same concepts. Thus, while a non-transitive notion of sameness is arguably enough to make sense of concept sharing, it does not suffice to guarantee that "thoughts are widely shared by chains of

relationist contender. But note that, as presented in the main text, the scope of the objection widens to any class of concepts one deems to be distinctively explanatory. As far as we can tell, the objection will then only be a 'small victory' for the non-relationist on the assumption that (i) indexicals express concepts which we may safely ignore in our account of concept sameness and (ii) that only indexicals express concepts with distinctive explanatory features. Neither of these claims seem however warranted.

communicating agents within and across linguistic communities, as well as chains of time slices connected by memory relations" (Onofri 2018, 19).

Regardless of the aforementioned points, some relationist authors have conceded that relational connectedness is *not* transitive (e.g. Fine 2007, Pinillos 2011, Prosser 2019, forthcoming). Prosser remarks that, just as Parfit's (1971) lesson about personal identity was that survival – understood as psychological connectedness – is not identity but a non-transitive relation between temporal stages, the lesson here might be that the same goes for the 'survival' of a concept (Prosser 2019, §7; see also Recanati 2016, Chap. 3). This is not the place to assess the cogency of Prosser's suggestive analogy. Instead, we will retain ourselves to defending the following two-fold claim: (i) a non-transitive notion of concept sameness compromises a substantive notion of concept publicity and (ii) relational theories appear obligated to admit that concept sameness is not transitive.

We take (i) to have been sufficiently established in the previous paragraphs. As for (ii): it is not hard to see that relational connectedness allow for cases in which a subject S_1 possesses two distinct concepts C_1 and C_2, the two of which are the same as a second subject S_2's C_3. This possibility remains open even if we assume that S_1 and S_2 represent (time-slices of) the same individual.

To illustrate, imagine the following scenario inspired by Kripke's (1979, 265-266) distinct – but not unrelated – discussion: Peter is exposed to two separate conversations about Paderewski, a famous pole pianist who has also acted as a prominent politician. In the first conversation, Paderewski is only referred to as being an accomplished pianist. In the second, as an important ex-prime minister of Poland. Given that Peter believes that no politician is a good musician, he infers that there must be two persons called 'Paderewski' being referred to in each conversation and thus entertains two distinct thoughts – one that he judges true, another that he judges false – which he would express by the same sentence, 'Paderewski has musical talent'.

Now, let us call Saul the person that was talking about Paderewski in the two contexts where Peter was present. We can assume that Saul knows very well that Paderewski, the pianist and politician, is no more than one person. That means that the concept that he was expressing in the two conversations by 'Paderewski' was one and the same. Let us call Saul's concept C_1. As we have seen, Peter's confusion led him to think that there were two distinct individuals named 'Paderewski'. In conceptual terms,

this means that Peter has at least two distinct concepts, C_2 and C_3, one of a polish musician, another of a polish politician. But by considerations of cognitive significance, C_2 and C_3 are surely distinct concepts, i.e. Peter can rationally take contrasting attitudes to two thoughts that differ only in the substitution of one concept for the other. However, it seems hard to deny that C_2 and C_3 are each relationally connected to Saul's C_1. For instance, it seems that Peter and Saul would trade on the identity of each other's tokens of 'Paderewski' in each of the two contexts where they were present and where a conversation about Paderewski took place. Similarly, it is plausible that if *a* forms a singular concept for some individual O in response to hearing *b* use a certain proper name N that refers to O, then *a* knows that the concept just formed co-refers with whichever concept *b* expressed by means of N. Given that principle, it is hard to deny that Peter knows that C_2 and C_3 co-refer with the concept that Saul expresses by means of 'Paderewski'. Indeed, provided that Peter formed C_2 and C_3 from listening to Saul talk about Paderewski, he would seem to have knowledge of co-reference between his concepts and whichever concept Saul expresses by 'Paderewski' – regardless of the fact that he is not in a position to infer that his two concepts, C_2 and C_3, also co-refer as a consequence of that.

The predicament should now be clear: C_2 and C_3 are (a) distinct from each other and (b) identical to C_1.[91] Cases with an analogous structure have led Prosser and others to give up on the transitivity of sharing a concept. As we have seen, however, it is not clear how a non-transitive account of concepts such as this would be able to account for a substantive notion of publicity.

Schroeter (2012) and Onofri (2018) are well aware that the non-transitivity menace hovers over their relational approaches. They suggest that, if a concept is shared between two individuals pertaining to the same community, then it should also be the case that this concept is shared by most members of that community. In order to make sense of this idea, they appeal to indirect linking relations and argue that two concepts or thoughts are the same, on their account, "not only when directly linked, but also when indirectly linked by a chain of direct linking relations" (Onofri 2018, 19). However, we

[91] Taschek (1998, 347 ff.) provides a variation of the point in the main text in terms of disagreement and agreement towards Peter's utterance of 'Paderewski has musical talent'. See also Fine 2007, Chap. 4.

submit, indirect linking relations do nothing to help us with the fact that Peter's two concepts are both linked with Saul's but nonetheless distinct from one another.

Cumming's (2013) view seems to be better geared to deal with Paderewski cases. According to Cumming, relational connectedness requires a one-to-one conventional mapping between concepts. To be sure, this is precisely what is not the case with respect to Peter and Saul – the former has two concepts mapping onto the latter's only one. Cumming is probably right that his formalization secures transitivity (symmetry and reflexivity) of the sameness relation (Cumming 2013, 14). However, even if he succeeds in this regard, Cumming's account is far too restrictive. It predicts failure of concept sharing in cases where it should not.[92]

Cumming focuses on cases where a subject takes two uses of the same name as being two uses of distinct homophone names. However, there are instances of cases like these in which it is extremely unlikely that one has ceased to share concepts with one's peers. Imagine the following twist to our story where the polish pianist himself goes to a classical concert and overhears some people saying, 'the pianist booked for tonight is not as good as Paderewski'. Given Paderewski's humbleness, he assumes that there must be someone (distinct from him) that shares his name and who is a good enough pianist to deserve such praise. As a matter of fact, that group consisted of Paderewski's closest friends and they were talking about him. However, given Paderewski's misapprehension, he now has two concepts that he would express by 'Paderewski' – one he knows is about himself and another he thinks is not. Provided that these two concepts are not one-to-one coordinated with his peers' only concept, they are also, on Cumming's account, not relationally connected with his peers' concept. Paderewski's uses of 'Paderewski', that is to say, will not express the same concept as his closest friends and relatives' uses of it. This is implausible. In conclusion, Cumming's proposal would only be able to save the transitivity of concept sharing at the cost of giving up on a plausible, less restricted notion of concept sharing, and hence again, a substantive notion of concept publicity.

[92] Onofri (2018, 18) makes a related argument.

3.2 Relational conditions are not necessary for publicity

So far we have been concerned with cases in which a relational condition between two concepts holds or plausibly holds, yet it is not thereby guaranteed that the concepts so related pass muster with central aspects of concept publicity, such as type classification, generalizable intentional explanation and transitivity. In this section, we purport to show that relational connectedness is not necessary for publicity either. Two (token) concepts that are not relationally connected may nevertheless be public and shared. To see this, we may reflect on two central phenomena: disagreement and basic concepts.

Disagreement. The difficulties for relational accounts vis-à-vis publicity also reach out into a domain that, at first pass, seems most amenable to them, namely, disagreement. We may focus on the most basic case in which subjects are in disagreement because they believe directly contradictory propositions of the form p and not-p. Under a widely shared conception of (dis)agreement, a necessary condition for two subjects S_1 and S_2 to be in (dis)agreement with each other about p is that S_1 and S_2 share the constituent concepts of p.[93]

The key idea is that disagreement in this basic sense is possibly an interaction-free notion. Relevantly for our purposes, this possibility is specifically attached to a conception of disagreement involving attitudinal states towards propositional contents (Cappelen and Hawthorne 2009, 60). Thus, MacFarlane observes that "[p]eople can be in disagreement even if they do not know of each other" (MacFarlane 2014, 119). Similarly, Teresa Marques notes that "we have the impression that there are disagreements between subjects who are not part of the same conversational setting, or do not even interact in any form" (Marques 2015, §1).

We may recur to several examples to prove the point. For instance, it is plausible that the average contemporary person disagrees with John Dalton over the indivisibility of atoms (Burge 1986, 716; Author paper 1). Arguably too, "the ancient Greeks were in disagreement with the ancient Indians about whether the bodies of the dead should be burned or buried even before Herodotus and other travellers made this disagreement known to them" (MacFarlane 2014, 119). But note, by contrast, that it does not seem

[93] Central cases of disagreement in this sense could also involve desires or intentions (e.g. Huvenes 2012, MacFarlane 2014, 122-123, Marques 2015). The same point holds, *mutatis mutandis*, for cases in which a subject changes her views across time.

correct to claim that the candidate relations advanced by Relationism actually hold between concepts whenever interaction-free disagreement holds. For instance, our atom concept and Dalton's do not seem to be relationally connected. In short, disagreement requires concept sharing but people may disagree even when no relational connection is present.[94]

To counter this line of argument, relationists may invoke again indirect linking relations between concepts. On this account, two people may disagree without being directly relationally connected as long as there is a possible chain of relational connections that eventually reaches out onto each other's concepts. However, indirect linking relations are here ineffective in a straightforward way. It is possible, and indeed conceivable, that two subjects S_1 and S_2 disagree in their beliefs over p while being neither directly nor indirectly linked. Consider, for instance, the inhabitants of Plato's mythical Atlantis and suppose that they actually existed but, limiting their whereabouts to the confines of their island, never really got in touch with any other community or civilisation. At one point, they disappear off the face of the Earth so that no trace of them is left. Let us suppose, finally, that Atlanteans had all sorts of controversial views over scientific, moral or religious issues. We may take these views to be incompatible with the views over the same matters of most other civilisations and communities in the rest of the world. It is thus conceivable that Atlanteans were in a wealth of disagreements with the rest of the world. However, *ex-hypothesis*, no relation and, *a fortiori*, no indirect linking relation yoked them to the rest of the world.[95]

A rather different strategy on behalf of Relationism would be to fall back on a purely truth-conditional account of disagreement. Even if ancient Greeks and Indians did not share their concepts, a relationist could say, they did have thoughts with the same truth-conditions. Our intuitions of disagreement between these two civilizations could then be accounted for by the fact that they often took contrasting attitudes to thoughts with the same truth-conditions.

---

[94] Although disagreement has attracted much more attention, the same point could be put in terms of agreement.

[95] The Atlantis case also blocks the suggestion that asymmetric or one-way concept relations may, as it were, trigger symmetric sameness concept relations (cf. Cumming's notion of symbol coordination that is symmetric but itself based upon the asymmetric notion of symbol mapping (2013, 6-12, esp. n. 32)). On this account, for C_1 to be the same as C_2, C_1 has to stand in R to C_2 but not the other way around. Irrespective of this, of course, Atlanteans are neither symmetrically nor asymmetrically related to any other community.

The issue here is that disagreement seems to be a hyperintensional notion, and a purely truth-conditional surrogate of it would not seem to be able to account for certain special situations. Imagine, for the sake of illustration, that both the ancient Greek and Indian civilizations used to observe the planet Venus in the mornings and in the evenings without knowing that it was Venus in both occasions. We could even suppose that, by sheer chance, they used the same names for Venus in these two occasions, 'Phosphorus' and 'Hesperus'. Now, suppose that the Greeks believed that Hesperus was a planet, but that Phosphorus was a star, whereas the Indians believed that both Hesperus and Phosphorus were planets. Intuitively, we would like a theory of concepts to allow us to say that these two civilizations agreed about Hesperus being a planet but disagreed about Phosphorus being so. On the other hand, since 'Hesperus' and 'Phosphorus' just refer to Venus, a purely truth-conditional conception of disagreement would not let us go so far: there is no truth-conditional (dis)agreement about Hesperus that would not immediately translate into a (dis)agreement about Phosphorus.

At this point, relationists could bite the bullet and admit that concept sharing stemming from disagreement relations is just out of scope for the view. But this would definitely not be a victory for Relationism. We are in this essay concerned with the question of whether Relationism has the resources to fully accommodate a substantive notion of publicity. It is in this context self-defeating to propose that Relationism does fail to explain disagreement-based concept shareability, or even worse, to retreat to a notion of disagreement without concept sharing, and hence without the publicity of our concepts.

Basic concepts. Leaving aside the issue of disagreement, it seems in general plausible that people can and do share concepts without (their tokens) standing in any form of relational connection to one another. A leaf from Fodor's writings is helpful here:

> [...] it should turn out that people who live in very different cultures and/or at very different times (me and Aristotle, for example) both have the concept FOOD; and that people who are possessed of very different amounts of mathematical sophistication (me and Einstein, for example) both have the concept TRIANGLE; and that people who have had very different kinds of

> learning experiences (me and Helen Keller, for example) both have the concept TREE; and that people with very different amounts of knowledge (me and a four-year-old, for example) both have the concept HOUSE. And so forth. Accordingly, if a theory or an experimental procedure distinguishes between my concept DOG and Aristotle's, or between my concept TRIANGLE and Einstein's, or between my concept TREE and Helen Keller's, etc. that is a very strong *prima facie* reason to doubt that the theory has got it right about concept individuation or that the experimental procedure is really a measure of concept possession. (Fodor 1998, 29)

We may put the point in terms of basic concepts. It is unlikely that each and every case in which it is plausible that two subjects – however spatio-temporally distant and culturally disparate – share one of these basic concepts is a case in which a relational condition is satisfied. People in different communities, speaking different languages or, in the extreme case, inhabiting different remote galaxies may not ever interact or possibly interact with one another. Still, it would be surprising if that would need to entail that these people never share any one concept – say, the concept of food, dog, or triangle, to use Fodor's examples or, perhaps even more plausibly, the logical concept of conjunction, or the material conditional. But this is what seems to be required on an account of sameness as relational connectedness. Note that the point holds water no matter what particular notion of basic concept we may be ready to countenance, so long as one is indeed allowed. If there are any basic concepts – be these, say, genetically, developmentally or rationally basic – people may plausibly share them without being relationally connected. And while relationists may try in this context to appeal to indirect linking relations or more complex construals of the target relations to accommodate this fact, they would, it seems to us, founder for exactly the same reasons we have already encountered.

One line of reaction to this point is to claim that Relationism's proper focus is not basic but non-basic concepts. While we lack a decisive argument against this manoeuvre, it does not however seem very advisable. First, if there is one explanatory target that our account of concept sameness should prioritise that is precisely basic concepts. If we

cannot get a grip on what concept sameness means for these concepts, we suggest, our account would be severely restricted. Second, to bite the bullet in regards to basic concepts seems, ultimately, to expose any relational account for a concept C to the criticism that C might actually be a basic concept. Basic concepts might plausibly include those for which Relationism is supposed to be giving the right sort of treatment, such as concepts expressible with perceptual demonstratives – such as 'this'/'that' or 'this F'/ 'that F' – and indexicals – such as 'I', 'here' or 'now'.

Another line of reply on behalf of relationists is perhaps to go modal: two concepts would be said to be the same just in case there is a possible world in which R holds between them. The problem with this suggestion is that the approach would border on idleness: arguably, for any two co-referring concepts C_1 and C_2 expressible with terms t and t' there will always be a possible world in which R holds. The proposal would therefore need to constrain the relevant modality in a way that does not result in a merely referential individuation of concepts, nor is blatantly *ad-hoc*. Pessimism about the prospects of such an endeavour would not seem unfounded.

4 Publicity and Reference-Determination

We do not take the lines of criticism set out in the previous section to severally constitute a knock-down argument against relational treatments of publicity. We do however believe that they collectively compound to make a strong case against such treatments. But the significance of this predicament would not be adequately portrayed without a diagnosis of the kind of hurdles we have been discussing. Here is what we take to be the key insight in this regard: both non-relational and relational approaches to concepts provide suitable accounts of one of the central basic roles that concepts play, namely, the CS-Role. However, when it comes to the the P-Role, neither theory seems to fully capture the phenomenon. On the one hand, non-relationists face the threat of postulating too demanding and variable concepts that are not likely to be shared from individual to individual. By contrast, Relationism puts forward a criterion of concept sameness that bypasses the variability issue but cannot account for the aspects of publicity brought out by the arguments outlined above, and thus ends up not guaranteeing a substantive notion of concept publicity.

An important point of this diagnosis is that non-relational theories are, perhaps surprisingly, largely untouched by the publicity problems we have been considering in the previous Section 2. To see this, we may reflect on traditional views as embodied in the idea of a 'referential theory'. In order to fix ideas and to offer the most encompassing version of the class of theories we are evoking, we may recur to the notion of a reference-determining condition. A reference-determining condition individuating a concept C is the condition – intensionally or extensionally described – that an object *x* has to fulfil in order for *x* to be the reference of C. When the condition for an object *x* to be the reference of C (say, *Hesperus*) is just that *x* be a particular referent *a* (say, Venus), then the reference-determining condition is 'purely' referential. Yet there is no need to subscribe to that particular interpretation. A condition for *x* to be the reference of C could be specified, intensionally, in terms of a mode of presentation or way of thinking taken as a primitive notion, a descriptive content (e.g. 'the heavenly body visible in the evening'), a judgemental disposition (e.g. application conditions for C) or a particular epistemic connection or knowledge towards *x* (e.g. acquaintance with or knowledge of Venus). This characterisation thus leaves considerable leeway as to the favoured details filling in the notion of a reference-determining condition.[96] What is relevant is that such a condition, however exactly glossed, is intrinsically attached to the concept in ways that serve, at the same time, to its individuation and to its reference-determination.

On this construal, referential theories capture the fact that two people share a particular concept in terms of these people using (token) concepts that share a reference-determining condition. For a concept to be public is, accordingly, for many people in the community to share (token) concepts with the same reference-determining condition. This broad-stroke framework can suffice to make vivid the way in which referential theories are supposed to help in relation to the aspects of publicity related above. For shareability, when understood in the referential terms just sketched, delivers the aspects of publicity we have been highlighting for free.

To begin with, a referential theory would afford the right results regarding scenarios that necessitate shareability but not external conceptual relations, namely, disagreement and basic concepts. In particular, interpersonal disagreement in different

[96] As the most elaborated non-relational accounts illustrate (e.g. Chalmers 2011, Evans 1982, García-Carpintero 2000, Jackson 1998 or Peacocke 2008).

communities and spatio-temporal locations may involve the same concepts so long as their reference-determining conditions are the same. Basic concepts also fit the mould. A referential theory would suggest that basic concepts are concepts individuated by basic reference or validity conditions cutting across relationally unconnected groups of thinkers.

Sufficiency problems are, if shareability is granted, clearly amenable to the referential framework too. Remarkably, the framework avoids any problems with transitivity: for two concepts to have the same reference-determining condition as a third concept, is for them to actually share such a condition. But a referential theory of the sort we are envisaging also fares better than Relationism when addressing the type classification and the intentional explanation worries. This theory would offer accounts in which relevant conceptual types correspond neatly with referentially articulated types and where the explanatory features engaged in intentional explanations is given by generalizable features of a reference-determining condition. Referential theories would then dodge the charge of being in principle blind to (semantic) conceptual type classifications or generalizable explanatory features.

The picture we arrive at is thus one in which interrelated weaknesses and strengths of referential and relational theories are brought to the fore vis-à-vis CS- and P-roles. In sum, for the reasons aptly pointed out in the relationist literature, referential theories account for the CS-Role in ways that generate trouble when it comes to guaranteeing shareability. Nonetheless, as just noted, referential views also put forward frameworks in which to assume shareability is quite nearly to assume publicity in the full sense we have been underlining. By contrast, relational theories do a good job securing concept shareability while providing insightful explanations of the CS-Role. However, relational shareability has trouble guaranteeing a notion of publicity covering the aspects under focus here. It seems therefore fair to conclude that, for different reasons, neither type of theory has so far offered an adequate account of the P-Role.

One may wonder: why should we care about the aspects of publicity that a referentialist theory seems so apt, and a relational theory so inapt, to tackle? Admittedly, the concerns of this paper carry weight only on the assumption that the aspects of publicity we have been highlighting are at all important. What if our theory only focuses on cases of direct interpersonal exchanges such as, paradigmatically, face-to-face

communication? In reaction to this we only want to note that it seems unassailable that one's theory would be better off if able to accommodate the full array of phenomena comprised under the idea that concepts are public. We take the ones we have tabled to be well-motivated on intuitive grounds. Whether or not our primary interest lies elsewhere, in short, it seems wise not to let those phenomena go unheeded.

5 Conclusion

In this article, we have argued that in spite of the many merits of Relationism, this recently developed family of theories fails to provide a full account of concept publicity. Reflection on the intuitively plausible manifestations of publicity brings out that relational connectedness is neither sufficient nor necessary for concept publicity. We take this insight to be an important corrective upon the hopes pinned on this kind of approaches but also, and perhaps more importantly, an invitation to take stock and reconsider the explanatory assets of non-relational views of concept sameness. According to these views, concept sharing is (not a mere stroke of luck but) a consequence of the fact that concepts are entities that make it possible for subjects to think of a world that is, for the most part, open and accessible to all. To individuate a concept is thus to single out a type of thinking about something in this open world. On this basic picture, to focus merely on external relations – as Relationism would have thus far suggested – will, very likely, only give you part of the true story about concept publicity. The suggestion is, therefore, that concepts can hardly be public if there is no intrinsic individuating relation between the concept and its publically accessible referent. Even if many, especially from relationist ranks, will find this suggestion unattractive or ultimately inviable, its serious consideration seems unavoidable in a fully satisfactory treatment – whether relational or not – of the idea that concepts signal, in Frege's words, a 'common stock of thoughts'.[97]

[97] Frege (1984, 185).

References

Campbell, John (1987) ‘Is Sense Transparent?’ *Proceedings of the Aristotelian Society* 88, 273-292.

Cappelen, Herman & Hawthorne, John (2009). Relativism and Monadic Truth. Oxford University Press UK.

Cappelen, Herman and Josh Dever (2013) The Inessential Indexical. Oxford: OUP.

Chalmers, David J. (2011) ‘Propositions and Attitude Ascriptions: A Fregean Account’, *Noûs* 45, 595-639.

Cumming, Samuel (2013) ‘From Coordination to Content’, *Philosophers’ Imprint* 13, 1-16.

Dickie, Imogen and Gurpreet Rattan (2010) ‘Sense, Communication and Rational Engagement’, *Dialectica* 64, 131–51.

Evans, Gareth (1982) The Varieties of Reference, Oxford: OUP.

Fine, Kit (2007) *Semantic Relationism*. Oxford: Wiley-Blackwell.

Fodor, Jerry A. (1998) Concepts. Where Cognitive Science Went Wrong. Oxford: Clarendon Press.

Frege, Gottlob (1984) ‘Concept and Object’, in McGuinness (ed.), M. Black et al. (trans.), Collected Papers on Mathematics, Logic, and Philosophy. Oxford, Blackwell, 182-194.

García-Carpintero, Manuel (2000) ‘A Pressupositional Account of Reference Fixing’, *The Journal of Philosophy* 97, 109-147.

Gray, Aidan (2017) ‘Relational Approaches to Frege’s Puzzle’, *Philosophy Compass* 12.

Heck, Richard G. Jr. (2012) ‘Solving Frege’s Puzzle’, *The Journal of Philosophy* 109, 132-174.

Heck, Richard G. Jr. (2014) ‘In Defense of Formal Relationism’, *Thought* 3, 243-250.

Huvenes, Torfinn Thomesen (2012) ‘Varieties of Disagreement and Predicates of Taste’, Australasian Journal of Philosophy 90, 167-181.

Jackson, Frank (1998) *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford University Press.

Kripke, Saul (1979) ‘A Puzzle about Belief’, in A. Margalit (ed.) *Meaning and Use*. Dordrecht: Reidel, 239-283.

Löhr, Guido. (2018) ‘Concepts and categorization: do philosophers and psychologists theorize about different things?’ *Synthese*. 10.1007/s11229-018-1798-4

Machery, Édouard. (2009). *Doing without concepts*. Oxford University Press.

MacFarlane, John (2014) *Assessment Sensitivity: Relative Truth and its Applications*. Oxford: Oxford University Press.

Maier, Emar (2015) ‘Reference, Binding, and Presupposition: Three Perspectives on the Semantics of Proper Names’, Erkenntnis 80, 313-333.

Marques, Teresa (2015) ‘Disagreeing in Context’, Frontiers in Psychology 6.

Ninan, Dilip (2016) ‘What is the Problem of De Se Attitudes?’, in S. Torre and M. García- Carpintero (eds.) *About Oneself*. Oxford: Oxford University Press.

Onofri, Andrea (2018) ‘The Publicity of Thought’, *The Philosophical Quarterly* 68, 521-541.

Parfit, Derek (1971) ‘Personal Identity’, *Philosophical Review*, 80: 3-27

Peacocke, Christopher (2008) *Truly Understood*. Oxford: Oxford University Press.

Perry, John (1979) ‘The Problem of the Essential Indexical’, *Noûs* 13, 3–21.

Perry, John (2006) ‘Stalnaker and Indexical Belief’, in J. Thomson and A. Byrne (eds.) *Content and Modality: Themes from the Philosophy of Robert Stalnaker*. Oxford: Clarendon Press, 204-221.

Pinillos, Nestor (2011) ‘Coreference and Meaning’, *Philosophical Studies* 154, 301-324.

Prosser, Simon (2015) ‘Why are Indexicals Essential?’ *Proceedings of the Aristotelian Society* 115, 211-233.

Prosser, Simon (2019) ‘Shared Modes of Presentation’, *Mind & Language*.

Prosser, Simon (forthcoming) ‘The Metaphysics of Mental Files’, *Philosophy and Phenomenological Research*.

Recanati, François (2016) *Mental Files in Flux*. Oxford: Oxford University Press.

Schroeter, Laura (2012) ‘Bootstrapping our Way to Samesaying’, *Synthese* 189, 177-197.

Taschek, William W. (1995) ‘Belief, Substitution, and Logical Structure’, *Noûs* 29, 71-95.

Taschek, William W. (1998) ‘On Abscribing Beliefs: Content in Context’, *The Journal of Philosophy* 95, 323-353.

SUMMARY

In this dissertation, I have investigated several philosophical puzzles associated to the thesis that thoughts are public, i.e. that in successful instances of communication, understanding, and in cases where thinkers are in genuine agreement with each other, the relevant thinkers accept the same thoughts. In chapter 1, I showed that this thesis seems difficult to uphold in the face of cases involving indexical expressions. When subjects successfully communicate with indexical expressions, they are nonetheless disposed to perform different actions, and thus we have reason to deny that they accept exactly the same thoughts. In chapter 2, I showed that this thesis is in conflict with the thesis that thoughts must track the cognitive profile of our attitudes ('Frege's Constraint'). In chapter 3, I showed that this thesis is in conflict with a minimal version of semantic internalism and that even the most conservative way of trying to make these two theses compatible involves weakening the claim that thought is public in the sense previously defined. In chapter 4, I investigated criteria of successful communication and argued against one based on match of referential content plus absence of false beliefs. In its place, I suggested we go back to criteria based on match of modes of presentation (thoughts) or successful recognition of the speaker's referential intentions. In chapter 5, I argued that thought's publicity cannot be fully accommodated by extant relationist theories of thoughts and concepts. One way to frame the most general conclusion of this dissertation is that it is futile to try to individuate an intersubjective notion of thought which is transitive, or which is equally useful from an intrapersonal perspective. If we have any reason for carving up an intersubjective notion of thought – and not even this is clear, as far as this dissertation is concerned – then it will most likely be orthogonal to the usual subjective one.