THEORIA

THEORIA, 2019
doi:10.1111/theo.12201

Communicating and Disagreeing with Distinct Concepts: A Defense of Semantic Internalism

by

MATHEUS VALENTE
Universitat de Barcelona

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's 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 article is preceded by a comparison of my view with another recently proposed by Recanati (section 7) and some replies to possible objections (section 8).

Keywords: semantic internalism, deference, concepts, thought, disagreement, communication

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.[1]

1 I will employ the convention of using words in capital letters to refer to concepts.

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 here are compatible with either 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 objectives of this article 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.[2] 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's (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 a few objections (section 8).

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.

2 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.

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 non-committal as possible.[3]

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 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.

3 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.

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.[4] However good its pedigree, it has come under serious attack in the latter half of the twentieth century from the so-called externalists.[5] 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 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 this article 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 of Internalism is not as convincing as some philosophers make it out to be and will argue that there are 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

4 As Johnston and Leslie (2012, p. 116) remark, "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".

5 Among those, Putnam (1975), Burge (1979) and Kripke (1980) stand out.

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 will 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 to classify 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 roles of concepts is to capture our most basic ways of keeping track of a topic in thought" (Schroeter and Schroeter, 2016, p. 5). It is not unusual to express this idea as a general constraint on a theory of concepts:[6]

(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.[7]

The most vivid way of arguing in favour of Publicity is by considering a story from Loar (1976, p. 357) 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

6 The name "publicity" comes from Onofri (2016).

7 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 are not other philosophically interesting cases of (dis)agreement. I will return to this in section 8.

(*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 in the opposite direction. I take it that most externalist arguments against Internalism are based on its conflict with Publicity.[8] 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.[9] 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.[10]

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

8 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.

9 One highly illustrative example is Chalmers's (2011a, pp. 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's project. In summary, I believe that knowledge of sameness of extension based on concepts' rules could help account for S-appropriateness in Chalmers's context.

10 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 nineteenth century, then it seems that the same concept has survived unscathed through major revisions in the rule that 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.

successful communication (genuine agreement) from rules. The conservative ways, in their turn, are those that weaken, but do not completely eliminate, the explanatory relation between them:

(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.[11] It is not my objective in this article 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 in more depth.

Conservativism about Internalism maintains the connection between concept expression and rules but weakens the extent to which one determines the other. One way of fleshing that idea out is by claiming that, contrary to Internalism, we

11 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 a 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.

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 thinkers' 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's (2001) defence 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's 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 favour 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 in meaning. (Wikforss, 2001, 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 arises. Similar stories abound in the medicine and clinical psychology literature. It seems plausible that psychologists these days have the same concept of autism that their early twentieth-century predecessors did, 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, 2018).

Now, everybody more or less already agrees about those points: concepts are as dynamic as the theories of which they are part. Indeed, many liberal philosophers (e.g., externalists) have used this type of consideration 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 in 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 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.

First, 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 section. Second, 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 that of both concepts, it is identical to the second's. I can think of two ways by which one could try to amend Approximation Internalism, but neither of them is successful.

The first would be to claim 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 neighbour, 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 every time we wrote or read "THARTHRITIS".

A second possible refinement of Wikforss's 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, 2001, 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.[12] Both SAI and NAI clearly fail.

In summary, there is one limitation and one problem with Wikforss's 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 weigh significantly

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

against the strategy of weakening Internalism in order to solve the dilemma that is the focus of this article. 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.

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 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 shy away 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 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. Somewhat 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 does not 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 section. Thus, the more we know about some subject matter, the less deferential our concept is and the more things we are able to do with it.

I think it is clear that 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 be helpful for 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 of 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 fulfil 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 inspection, 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 favour.

The failure of the first two options leaves 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 communicate successfully 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 exactly the same concepts as long as there is some special relation holding 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 genuine agreement 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 is precisely what we have seen when confronted with Loar's story back in section 2. Let us revisit this. 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 communicate successfully 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.[13] Here is one model of

13 Sameness of extension that can be known on the basis of facts that are extrinsic to the concepts' rules or to the representations that 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.

how that could happen: if it is 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.

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 is a unique F-and-G).[14]

Let us 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.

14 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, 2016, pp. 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.

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 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 communicate successfully (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 communicate successfully 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 (2016, ch. 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,[15] 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, while Recanati talks of fine-grained and coarse-grained concepts,[16] 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 (especially with such complicated terms-of-art such as "concept", whose meanings are constantly up for

15 In Recanati's terminology, rules are epistemically-rewarding relations (see Recanati, 2012).
16 Actually, Recanati talks of "static mental files" and "dynamic mental files".

grabs). However, I think there is at least one point in favour 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.[17]

Another difference is that Recanati (2016a, pp. 71–94) 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. 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.[18] Thus, communication with distinct concepts ends up sounding like a risky endeavour.

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.

---

17 Schroeter and 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 and 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.

18 Recanati admits of only one possible type of exception but relegates it to a footnote (2016a, p. 94, n. 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.

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 (2016, pp. 89–94) argues convincingly that memories additionally locate the source of their originating perceptual experience in the 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.[19] 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.

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, pp. 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

19 As Recanati (2016, 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.

is said – even if, given their distinct contextual stipulations, their "tall" tokens apply to distinct people (e.g., according to A, people above 1.80 m count as tall; according to B, people above 1.90 m). The concepts A and B express by "tall" do not necessarily have the same extension. Indeed, if Rachmaninoff's height were 1.85 m, 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 should not 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.[20] One could say, for example, that our intuition is based on the contingent fact that Rachmaninoff 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*.[21] Each strategy will have virtues and defects that I will not discuss, 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 neighbourhood of other objects, and D disagrees, should we not characterize C and D as being in genuine disagreement regardless of the fact that the concepts they express by "planet" are not only distinct but also

20 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; Valente, 2018).

21 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.

pick out distinct things? The reply here, as in the previous case, is that the type of disagreement between C and D does not 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 and Sundell, 2013), and not as in genuine disagreement with respect to the contents they express.[22]

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. As one example, consider the following case.[23] In a recent experiment, half of the Dutch participants who were asked to colour the part of a drawing of a human body corresponding to the arm (in Dutch, "arm"), coloured the drawing from the shoulder to the wrist, while the other half coloured 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".[24]

In any case, I agree with Plunkett and 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.

---

22 Plunkett and 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.

23 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.

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

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 travelling in a foreign country whose native language she knows nothing about; she can, however, on every occasion on 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's 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 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 communicate successfully by it. That is not a desirable consequence; it seems undeniable that Dolores's metalinguistic trick should not 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's). 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 does one need 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 here help views like Internalism to gain momentum.

Acknowledgements

The material on which this article is based was presented in Barcelona, both at the Metaphysics Seminar and at the Diaphora Midterm Conference in 2017, and in Porto, at the II Mind, Language, and Action Group Graduate Conference in 2018. I thank the audiences at these three occasions, as well as several colleagues who constantly provided me with valuable feedback, and two reviewers for *Theoria* for their important suggestions.

References

Braddon-Mitchell, D. (2004) "Masters of our Meanings." *Philosophical Studies* 118(1–2): 133–152.

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

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

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

Chalmers, D. 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. (2014) "Troubles for Content I." In A. Burgess and B. Sherman (eds), *Metasemantics: New Essays on the Foundations of Meaning*, pp. 147–168. Oxford: Oxford University Press.

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

Johnston, M. and Leslie, S. (2012) "Concepts, Analysis, Generics and the Canberra Plan." *Philosophical Perspectives* 26(1): 113–171.

Kripke, S. A. (1980) *Naming and Necessity*. Harvard University Press.

Loar, B. (1976) "The Semantics of Singular Terms." *Philosophical Studies* 30(6): 353–377.

Majeed, R. (2018) "Why the Canberra Plan Won't Help You Do Serious Metaphysics." *Synthese* 195: 4865–4882.

Majid, A. (2010) "Words for Body Parts." In B. C. Malt and P. Wolff (eds), *Words and the Mind. How Words Capture Human Experience*, pp. 58–71. Oxford: Oxford University Press.

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 S. Torre and M. García-Carpintero (eds), *About Oneself: De Se Thought and Communication*, pp. 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." In T. Marques and Å. Wikforss (eds), *Shifting Concepts*. Oxford: Oxford University Press.

PLUNKETT, D. and SUNDELL, T. (2013) "Disagreement and the Semantics of Normative and Evaluative Terms." *Philosophers' Imprint* 13(23): 1–37.

PUTNAM, H. (1975) "The Meaning of 'Meaning'." In H. Putnam (ed.), *Mind, Language and Reality: Philosophical Papers*, Vol. 2, pp. 215–271. Cambridge: Cambridge University Press.

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

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

SCHROETER, L. and SCHROETER, F. (2016) "Semantic Deference versus Semantic Coordination." *American Philosophical Quarterly* 53(2): 193–210.

SHOEMAKER, S. (1970) "Persons and Their Pasts." *American Philosophical Quarterly* 7: 269–285.

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): 217–231.