The Epistemology of Immunity to Error through Misidentification

This paper offers several new insights into the epistemology of immunity to error through misidentification, by refining Pryor’s (1999) distinction between de re misidentification and wh-misidentification. Such a refinement is crucial for identifying exactly what is at issue in debates over the Immunity thesis that, roughly, all introspection-based beliefs about one’s own occurrent psychological states are immune to error through misidentification. I contend that the debate between Campbell (1999) and Coliva (2002) over whether the phenomenon of thought insertion provides empirical evidence against claims like Immunity has wrongly focused on de re misidentification as the relevant sort of misidentification in question, and largely overlooked the role of wh-misidentification. I argue that, once we properly distinguish the two notions, subjects of thought insertion can be seen to make an error of wh-misidentification in their judgments. I argue that this disproves the Immunity thesis, so properly understood, and show what broader implications this has for the epistemology of IEM as well as our understanding of the first person.

Philosophers since Wittgenstein (1958) and Shoemaker (1968) have observed that many of our first-personal judgments based in introspection appear to exhibit immunity to error through misidentification (or IEM). This is, roughly, the idea that when I introspect and thereby come to judge whether I am F, I may be wrong about whether I really am F, but I cannot be wrong about who is F; I am F, if anyone is. This is thought to be a general feature of our basic introspective judgments. As a first pass principle:

**Immunity**

Any introspection-based belief about one’s own occurrent psychological states is immune to error through misidentification

Many have upheld the claim of Immunity, under some qualified form or another, as a conceptual truth (see Wittgenstein 1958, Anscombe 1975, Peacocke 1999, Shoemaker 1968, Evans 1982, Wright 2012). Nonetheless, there are reasons to dispute its status as an *a priori* truth. In the Campbell-Coliva debate, Immunity has been challenged on empirical grounds: schizophrenic patients who are victim to so-called thought insertion are able to form what appear to be introspection-based beliefs that are liable to error
through misidentification.\textsuperscript{1} Campbell (1999a, 2002) believes the phenomenon of thought insertion poses a genuine counterexample to Immunity; Coliva (2002a, 2002b) denies this.\textsuperscript{2} Problematically, however, little attempt has been made to define the relevant notion of IEM in precise terms. For there is more than one type of misidentification, and it is controversial which of these applies to Immunity. Insofar as attempts have been made, those within the Campbell-Coliva debate have tended to focus on just one type of misidentification—\textit{de re} misidentification.\textsuperscript{3} This tendency, I shall argue, is unjustified.

This paper sets out to clarify several issues concerning the epistemology of immunity to error through misidentification, by improving upon earlier insights had by previous accounts. Doing so, I shall argue, reinstates the objection from thought insertion against Immunity. This partially amends the situation of such debates over Immunity, by providing a rigorous treatment of all the various possible sorts of misidentification and identifying exactly which of these is under dispute.

I first introduce and offer a refined version of Pryor's (1999) distinction between \textit{de re} misidentification and \textit{wh}-misidentification (§§I–II). My presentation brings out several features missing in Pryor's account. These will be necessary for putting the theory of IEM to use in assessing the Immunity debate. I contend that the Campbell-Coliva debate over Immunity has wrongly focused on \textit{de re} misidentification as the relevant sort of misidentification in question. Rather, it is \textit{wh}-misidentification that is at issue.\textsuperscript{4} I shall argue that thought insertion refutes the Immunity thesis, so understood—namely, that all introspection-based present-tense psychological self-ascriptions are immune to error through \textit{wh-}

\textsuperscript{1} See also Lane and Liang (2011) and Vignemont (2012), who both offer somatoparaphrenia—the bodily corollary of thought insertion—as empirical evidence against the ‘bodily’ variant of Immunity.

\textsuperscript{2} For other defenses of Immunity, see also Gallagher (2000), Graham and Stephens (2000), Billon (2011).

\textsuperscript{3} Coliva, for instance, has only \textit{de re} misidentification (see §1 below) in mind when she claims that any purported counterexample to Immunity must “involve a belief in an identification component” (2002a:30).

\textsuperscript{4} This strategy of refocusing the debate in the IEM literature is not unprecedented. Pryor reinterpreted the Shoemaker-Evans debate to be over the alleged \textit{wh}-IEM status of memory self-ascriptions, rather than their alleged \textit{de re} IEM status, ultimately siding with Shoemaker (see Shoemaker (1968), Evans (1982), Pryor (1999)). This paper aims to reinterpret the Campbell-Coliva debate to be over the alleged \textit{wh}-IEM status of present-tense psychological self-ascriptions, rather than their alleged \textit{de re} IEM status, ultimately siding with Campbell. Relocating the issue in this way arguably advances the debate. Or so I suspect, since I am inclined to agree with Pryor (1999:286) that, between the two, \textit{wh}-IEM is the more basic and interesting type of IEM.
misidentification (§III). It need not, however, be seen as demonstrating *de re* misidentification (§IV). Nor does it essentially involve any kind of reference failure or other semantic defect (§V).

The ensuing proposal, if successful, shows why IEM should be construed as an epistemic rather than semantic phenomenon; why *wh*-misidentification remains, as Pryor believed, more interesting a notion than *de re* misidentification; and why the general phenomenon of IEM nonetheless offers limited hope of enriching our understanding of the first person.

**I. Immunity to error through *de re* misidentification**

Following Pryor (1999), I shall distinguish *de re* misidentification and *wh*-misidentification, and their respective forms of immunity. My account follows Pryor’s in broad outline, but differs in crucial ways. The details will matter when deciding whether or not to classify thought insertion as a case of misidentification. Let us begin with *de re* misidentification.

On one understanding, misidentification occurs when you base your belief on a mistaken identity assumption. To illustrate:

*Example 1.* Suppose you are looking for your friend Daniel at a ball game. You see a man among the bleachers waving at your direction. You believe he has reserved seats for you, and so, pointing to him, tell your other friend, “Look! Daniel’s waving.” But that man, unbeknownst to you, is not Daniel. You’ve mistaken someone else for Daniel.

This sort of mistake, known as *de re misidentification*, can be characterized precisely in epistemic terms:

**De re misidentification:** One’s grounds for his *de re* belief about an object *a* to the effect that it is *F* rests on his grounds for believing, of some *x*, and of *a*, that *x* is *F* and that *x* is identical to *a*; but unbeknownst to him, *x* ≠ *a*.5

---

5 This is a simplification of Pryor’s account, but will do for our purposes. Pryor adds two stipulations: (i) all the beliefs in question (that *x* is *F*, that *x* = *a*, that *a* is *F*) are of singular propositions, where a singular proposition about *x* to the effect that it is *F* is a proposition whose truth-value with respect to any possible world *w* depends solely on whether or not *x* is *F* in *w* and any belief in that proposition is a *de re* belief about *x*; and (ii) the target belief about *a*,
In the example, the grounds for your belief that Daniel is waving rest on your grounds for believing that That man is waving and that That man is Daniel, where the latter is a mistaken identity assumption. You may be right in thinking that that man is waving. But you’re wrong to think it is Daniel who is waving.

Judgments that exhibit immunity to error through misidentification in this sense are sealed off from the sort of mistake in Example 1. Let us say that a belief about an object a (under that very mode of presentation) to the effect that it is F is immune to error through de re misidentification (or exhibits de re IEM) just when it is a de re belief about a but its grounds do not rest on any non-trivial identity assumption involving a. A belief is liable (or immune) to error through de re misidentification not as such, but only relative to some justification or grounds for forming or holding that belief. Although any to the effect that it is F, is something the subject may either believe or simply aims to express as part of his basic referential attempt (which, in Example 1, would be a proposition about Daniel, since you attempt to refer to Daniel by referring to the man in the bleachers, and not the other way around: your first referential attempt is more basic than the second). See Pryor 1999: 273–5; n.17.

These extra stipulations arguably fail as necessary conditions. Given constraint (i), the account rules out target beliefs in non-singular propositions (e.g. if in the example I said instead, “My friend from Alabama is waving”, thereby aiming to say something about Daniel, whom I know is my only friend from Alabama), when intuitively these should count as exhibiting de re misidentification too. Given constraint (ii), the account also fails to generalize to cases of intended misdescription, where the target proposition is neither believed nor part of what the subject attempts to express (e.g. if I were to say sarcastically, “The only one rooting for the winning team on this side of the bleachers is waving”, and intend my remark to be understood as sarcasm—i.e. I neither believe nor intend to express, as part of what I mean, that Auburn will be the winning team). To be sure, Pryor offers no definition of de re misidentification, only a stipulative set of sufficient conditions. However I doubt that either (i) or (ii) should enter into a full account of de re misidentification, and so ignore them in what follows.

6 “Nontrivial”, so as to rule out unchallengeable identity assumptions, such as tautological self-identities of the form a = a or trivial existence claims of the form ∃x (x = a). Otherwise, virtually every judgment would be trivially liable to error through de re misidentification. Alternatively, one can constrain the relevant notion of identification, which, in the de re sense, really amounts to the notion of reidentification (which explains why one must be in a position to have de re beliefs about both the identified referent x and the target referent a).

Other related caveats are also absent from Pryor’s (1999: 279) definition. Pryor often talks as if one must have some justification or another for believing that x = a in order to de re misidentify x as a. Clearly, this is inessential: de re misidentification may occur even when the identity assumption is groundless—one simply assumes, without justification, that x is a. Pryor’s definition of de re IEM (of a proposition relative to some grounds of justification) as the impossibility of de re misidentification (when believing that proposition on those grounds) also risks over-classifying certain cases. A judgment that Hesperus is bright, based on believing that Phosphorus is bright and that Hesperus is Phosphorus, can never result in de re misidentification, so defined, insofar as the latter identity claim is necessarily true. Yet intuitively, de re misidentification is still possible in some sense—only here, the intended possibility is epistemic, not metaphysical. Letting ‘a’ and ‘x’ range over modes of presentation forestalls this worry.

7 We could, by extension, define a belief to be absolutely immune to de re misidentification just in case it is immune to de re misidentification when justified by every possible ground for believing it. However, I think Pryor is right to suspect that such a notion is empty. Even for present-tense psychological self-ascriptions that are usually justified on the basis of cogito-type reasoning—the standard paradigm of IEM—it is nevertheless possible for me to
instance of de re misidentification must involve a mistaken identity assumption, this assumption need not figure explicitly as a premise for inferring the target belief. You may form the belief that Daniel is there in the bleachers quite immediately and unreflectively, without having to consciously identify that man as Daniel; you simply (albeit wrongly) take that man to be Daniel.

Two observations deserve comment. First, in cases of potential de re misidentification, where there is predication of Fness involved, even if one questions his grounds for believing the identity assumption that $x = a$, he nonetheless retains his grounds for believing, of $x$, that it is F. There is, as it were, a “remainder” for de re belief when one commits (or suspects) an error of de re misidentification. Call this the de re remainder. Suppose you discover that the man in the bleachers is not Daniel after all. (You realize, to your embarrassment, that he is really waving at someone behind you, not you.) Despite being mistaken in thinking that man is Daniel, you still have grounds for believing, of that man, that he is there.

Second, there are two different ways of de re misidentifying any given object. One can either misidentify something else as the object or misidentify it as something else. Consider:

Example 2. Helga is the village’s notoriously ugly stepsister. I have never seen her before, but only caught word that Helga is ugly (i.e. have testimonial grounds to believe $x$ is $F$). One day I see a hooded stranger whom I take to be Helga (i.e. assume $x = a$), so I mutter, “She is ugly” (i.e. believe $a$ is $F$). Unbeknownst to me, she is not Helga.

believe these on the basis of non-introspective grounds and false self-identifications (e.g. believing that $I$ am $F$ on the basis of believing that Napoleon is $F$ and that I am Napoleon). See Examples 4–5 below and Pryor 1999: n.19.

8 “Potential de re misidentification”, because grounds for identity assumptions may be questioned, even if the identity claim in question is true as a matter of metaphysical necessity (so that no actual de re misidentification can occur). “Where there is predication of Fness involved”, because bare judgments of identity that are not conceptual truths, like “Hesperus is Phosphorus”, are trivially open to (the epistemic possibility of) de re misidentification, yet admit of no possibility of a remainder belief (since the target belief here just is the identity assumption in doubt).

9 In Prosser’s (2012) terminology: One can be mistaken in identifying the “source object” $x$ with the “target object” $a$, yet nonetheless justified in holding onto the de re belief about the source object $x$.

10 This is essentially Pryor’s (1999: 275) pair of examples of misidentifying the blue-coated man as Sam vs. misidentifying Sam as the blue-coated man.
Example 3. In a similar scenario, Helga’s notoriety does not derive from her ugliness. I have heard of Helga the stepsister, but have no reason to think she is ugly. One day I see a stranger, whose ugliness is now manifest (x is F). Mistaking her for Helga (x = a), I conclude, “Helga is ugly” (a is F).

In each case, I falsely assume that the stranger is Helga, only the original “predicative” grounds—for what I remain justified in believing, should that identity assumption come into question—is different, whether concerning Helga (Example 2) or the stranger (Example 3).

Both these features—the availability of a de re remainder belief and room for misidentifying in two different ways—carry over into cases of de re misidentification involving the first person. I can either misidentify someone else as myself (Example 4) or misidentify myself as someone else (Example 5).

Example 4: While working out, I look up and mistake someone else’s arm in the mirror to be my own. In a brief moment of satisfaction, I tell myself, “I flex well.”

Example 5. Suppose I lift weights while deluded: I believe I am Arnold Schwarzeneger. Based on how my muscles feel, without looking up, I tell myself, “Arnold flexes well.”

Both cases involve mistaken self-identification. In Example 4, I wrongly assume it is my own arm I see in the mirror. But even after I am disillusioned of my flex-ability, I still have grounds for believing, of the other man, that he flexes well. In Example 5, I mistake myself to be Arnold Schwarzeneger. My grounds for believing Arnold flexes well are undercut once I realize (after therapy) that I am not Arnold, though I still retain grounds for believing about myself that I flex well.

---

11 Pryor (1999: n.22) insists that describing things this way would make this a case of de re misidentifying my arm, but not de re misidentifying myself; since the false identity assumption involved (strictly speaking) concerns my arm, not me. I am not so convinced. On at least one reasonable interpretation, the case demonstrates what Evans (1982:199), following Quine (1968:194), calls “deferred ostension”: although the only object directly available to perception is (the reflected image of) some person’s arm, I can still use a complex demonstrative like “that guy” [whose arm I see in the mirror] to refer to, and entertain de re beliefs about, that person (or his entire body), and not just his arm. In that case, my grounds for believing I flex well rest on these perceptual grounds for believing That guy flexes well together with the false identity assumption I am that guy. This is evidenced by the fact that, even if the latter identity assumption comes into doubt, my immediate perceptual experience, though limited, nonetheless continues to afford me justification for holding the former de re belief about the “whole” guy (and not just his arm).
We shall revisit *de re* misidentification later when considering the alleged IEM of our psychological self-ascriptions.  For now, we turn to another type of mistaken identification, *wh*-misidentification.

II. Immunity to error through *wh*-misidentification

In the *de re* sense, misidentification results from a false identity assumption. On another understanding, misidentification is simply failing to correctly figure out which thing is *F*, but not necessarily because of any mistaken identity assumption.

To use Pryor’s (1999: 281) original example:

*Example 6.* Suppose I smell a skunky odor, and see several animals rummaging around in my garden. None of them has the characteristic white stripes of a skunk, but I believe that some skunks lack these stripes. Approaching closer and sniffing, I form the belief, of the smallest of these animals, that it is a skunk in my garden. This belief is mistaken. There are several skunks in my garden, but none of them is the small animal I see.

Here I fail to correctly figure out which thing it is that (or number of things each of which) is a skunk. I misidentify the source of the smells as the smallest of the animals in view. But, unlike with *de re* misidentification, here I make no mistaken identity assumption. It is not as if I had already previously identified the source of the skunky odor and am now reidentifying *it* as the smallest among the animals I currently see. Rather, I fail to correctly identify the object in the first place.

Call this *which-object misidentification* (or *wh*-misidentification), defined as such.  

---

12 Arguably many introspection-based present-tense psychological self-ascriptions exhibit *de re* IEM. Therefore, they leave behind no *de re* remainder when one’s grounds for identification are undercut, nor do they offer multiple ways of misidentifying the object in question, precisely because in each case there is no (non-trivial) identity assumption to be made. I postpone the question of whether all such beliefs do in fact exhibit *de re* IEM (see §4).

13 So called, since to *wh*-misidentify the *F*(s) is to incorrectly answer the question “Which thing(s) is (are) *F*?” Again, this is a simplified version of Pryor’s definition, which states that *wh*-misidentification occurs just when: (i) a subject has some grounds *G* that offer him knowledge of the existential generalization $\exists x\ Fx$ (in the sense that, if he were to believe it on the basis of *G*, and possessed no defeating evidence, then his belief would constitute knowledge that $\exists x\ Fx$); (ii) partly on the basis of *G*, the subject is also justified, or takes himself to be justified, in believing of
**Which-object misidentification:** One’s grounds for believing of some object $a$ that it is $F$ (is not $F$) rest in part on, but are not limited to, his grounds for believing that something is $F$, but unbeknownst to him, $a$ is not $F$ (is $F$).

(Ignore the parenthetical clauses for now; their significance will emerge later.) According to Pryor, my grounds for the target belief that *the smallest of these animals* is a skunk rest in part on, yet exceed, my grounds for the existential belief that *something* is a skunk. For I would cease to be justified in believing the first thing if my reasons for believing the second were discredited. (Perhaps I hear word of a neighborhood prank and suspect the skunky odor is completely artificial.) Yet what I smell, or whatever else justifies the existential belief, cannot exhaust my entire grounds for the target belief: there is a gap between believing that *something* is $F$ and believing that $a$ is $F$. Smelly experiences may allow me to know that a skunk is nearby, while leaving it open as to whether *the smallest of these animals* is a skunk.

The possibility of a remainder belief thus arises with *wh*-misidentification too, just as with *de re* misidentification; only here, it is the bare belief that Fness is being instantiated by some thing or another. Suppose I discover that none of the animals I see in my garden are skunks. I can no longer believe that the smallest *among them* is a skunk. Yet, supposing I am certain that the skunky smell really is from a skunk, I still have grounds for thinking nonetheless that *something* is a skunk nearby. Call this the *existential remainder*. In general, in cases of potential *wh*-misidentification, one’s grounds for believing that $a$ is $F$ may be undercut without undercutting his original grounds for believing that *something* is $F$.\(^1\)

Judgments insusceptible to misidentification of the *which-object* sort allow for no such gap between the existential remainder and the whole target belief. Let us say that a belief about $a$ to the effect that it is

\[
\text{some object } a \text{ that } \text{if } a \text{ is } F, \text{ instead some distinct object (or objects) } y \text{ is } F, \text{ and it’s because the grounds } G \text{ “derive” in the right way from this fact } \text{about } y \text{ that they offer the subject knowledge that } \exists x \text{ Fx. Pryor’s definition also lacks—as we shall see, crucially—the parenthetical clauses. See Pryor 1999: 281–2.}
\]

\(^{14}\) “Potential *wh*-misidentification” in an epistemic sense, because one’s grounds for believing $a$ is $F$ may be questioned, even if $a$ is $F$ is a necessary truth, so that no actual *wh*-misidentification can occur. When calculating a restaurant check, I may upon hurried calculation believe of some $n$ that it is 120% of the subtotal $m$, but suspend my belief should I question the accuracy of my mental math (“It can’t be *that* much!”), even if I did happen to calculate correctly (such that it is necessarily true of $n$ and of $m$ that $n = 1.2m$). Notice: unlike with *de re* misidentification, no qualification about “involving predication of Fness” is needed here, since no identity assumptions are involved.
F (or is not F) is **immune to error through wh-misidentification** (or exhibits **wh-IEM**) just when the subject’s grounds for believing that *something* is F do not leave it open to him as to whether *a* is F, but offer knowledge that *a* is F (or is not F). If so, any defeater that undercut one’s grounds for believing that *a* is F (or not F) would thereby undercut one’s original grounds for believing that anything is F.

There are several interesting yet potentially confusing ways in which Pryor’s example differs from standard cases of *de re* misidentification like Examples 1–5. First, the example involves a mistake on my part in a task of cross-modal binding: I wrongly bundle my olfactory information (the skunky odor I smell) with my visual information (the animals I see) and attribute both to one and the same object. Second, my visual experience is of a whole collection of animals rather than of a single object. Third, I am not in a position to hold a *de re* belief about the animal I’m smelling; I only know that something or other smells like a skunk and that among the whole bunch of animals I see there is one that is the smallest within that bunch. Fourth, there is no identity assumption involved in my attempt to identify the source of the smell.

The first two differences appear to be mere accidental features of the example as it is described and that the key differences lie in the last two respects. However else they may differ, wh-misidentification importantly allows for two things which *de re* misidentification cannot allow for: the absence of any *de re* belief and the absence of any identity assumption. Thus there are not just one but *two* types of wh-misidentification to distinguish apart from *de re* misidentification.\(^\text{16}\) Consider:

\(^{\text{15}}\) More precisely: a belief about *a* to the effect that it is F (or is not F) is **immune to error through wh-misidentification** (or exhibits **wh-IEM**) when **justified by grounds G** just in case for any part G* of the subject’s entire grounds G, if G* were to offer him knowledge that *something* is F, it would offer knowledge that *a* is F (or is not F). Alternatively: a belief of *a* that it is F (or is not F) is **vulnerable to wh-misidentification when justified by grounds G** just in case it is possible for grounds G to be defeated by undercutting evidence E such that their combination E+G no longer justifies the subject in believing that *a* is F (is not F) but E+G could by itself offer him knowledge that \(\exists x \text{Fx}\). See Pryor 1999: 283–4; n.28.

\(^{\text{16}}\) Now, while Pryor doesn’t run these two together, he doesn’t make any explicit note of their differences either.
Example 7. Suppose we are hiking up a snowy mountain. I find a trail of giant footprints in the snow, call them to your attention, and say, “The most ferocious living creature created these tracks.”

Here I am not in a position to hold any de re belief about whatever is actually the most ferocious living creature. I also do not base my judgment on any identity assumption. I merely take the most ferocious living creature to be whatever created those footprints. I may be wrong about that and guilty of wh-misidentification, but I would still be entitled to make the existential claim that something created these tracks. Suppose you tell me, “No, you’re wrong. The most ferocious living creature lives hundreds of miles away. I’m married to her.” Confronted with this defeating evidence, I could no longer justifiably say what I did, but I would still have grounds for saying the tracks were created by something.

Now consider:

Example 8. Suppose again that we are hiking and I find a trail of giant footprints. A while later, we encounter for the first time an animal with large feet. I call them to your attention and say, “That creature created those tracks.”

Unlike the previous case, here I am in a position to hold a de re belief about the creature. However, my target belief is still not based upon any identity assumption, for I am not reidentifying anything I had previously seen or come into contact with before. I simply identify the large-footed animal as the thing responsible for making the footprints. Again, it is possible for my grounds for holding that belief to be undercut while leaving an existential remainder (as when, suppose, we walk a bit further and see the tracks ending at the feet of another animal).

This differentiates two possible types of wh-misidentification: cases where both de re beliefs and identity assumptions are absent (Examples 6–7) and cases where only the identity assumption is absent (Example 8). There are also cases where both a de re belief and an identity assumption are involved. But
this is just \textit{de re} misidentification (Examples 1–5).\footnote{Are there cases of an identity assumption but no \textit{de re} belief? This, I take it, is impossible. Making an identity assumption involves reidentifying some object I have previously come into contact with. But I cannot reidentify $x$ as $a$ unless I am in a position to have \textit{de re} beliefs about both (e.g. previously seen) $x$ and (e.g. presently seen) $a$.} Notice that they count as \textit{wh}-misidentification too, since any undercutting defeater to one’s identity assumptions that leaves behind a \textit{de re} remainder ($x$ is $F$) thereby also leaves an existential remainder ($\exists x\,Fx$).\footnote{More carefully: Suppose one’s belief about $a$ that it is $F$, held on grounds $G$, rests on a false identity assumption between $x$ and $a$. Undercutting evidence $E$ may defeat one’s grounds for the identity assumption and hence the target belief. But the remaining grounds $E+G$ will still justify believing of $x$ that it is $F$. Hence it will still justify believing that \textit{something} is $F$, since the \textit{de re} remainder entails the existential remainder. So there is a part of the original grounds—namely, $E+G$—that offers knowledge of the existential, but not of the target proposition $a$ is $F$.} In that case, \textit{de re} misidentification is just another type of \textit{wh}-misidentification. It is \textit{not}, as Pryor and others have suggested, a separate type of misidentification, but merely a special case (but, as Pryor’s discovery showed, not the only case either).\footnote{I am unsure whether Pryor thinks the two are separate \textit{sui generis} categories. In certain places it seems he does: “\textit{Wh}-misidentification can occur without \textit{de re} misidentification … And \textit{de re} misidentification can occur without \textit{wh}-misidentification.” Yet elsewhere, he says: “if a belief is immune to \textit{wh}-misidentification when justified by certain grounds, that entails it is also immune to \textit{de re} misidentification when justified by those grounds. (Contraposing, any beliefs which are vulnerable to \textit{de re} misidentification when justified by certain grounds are ipso facto also vulnerable to \textit{wh}-misidentification when justified by those grounds. )” (1999: 285).} Consequently, any belief that is immune to error through \textit{wh}-misidentification when held on certain grounds is thereby immune to error through \textit{de re} misidentification when held on those same grounds. Examples 6–8 show the converse fails. In sum: \textit{de re} misidentification implies \textit{wh}-misidentification (but not vice versa) and \textit{wh}-IEM implies \textit{de re} IEM (but not vice versa).\footnote{One may object: “Isn’t \textit{wh}-misidentifying something as the $F$ just attributing $F$ness to the wrong thing? In that case, \textit{wh}-misidentification is not a genuine form of misidentefication at all. It is simply a form of mistaken predication of $F$ness. Indeed, it looks as if any error of judgment will count as \textit{wh}-misidentification: mispredicating $F$ness of something is failing to figure out which thing (if any) is $F$. This trivializes your result that \textit{de re} misidentification is a proper subcategory of \textit{wh}-misidentification, because a mistaken identification is trivially one way among many of making a mistaken judgment. Equivalently, it is trivially true that \textit{wh}-IEM implies \textit{de re} IEM (but not vice versa), because any incorrigible judgment will trivially be immune to all errors, including errors of mistaken identifications (but not vice versa).” To be sure, \textit{wh}-misidentification is \textit{a form} of mispredication, as opposed to an error of making a false identity assumption. But this is the very point of distinguishing \textit{wh-} from \textit{de re} misidentification: misidentification need not involve identity mistakes. However, it is incorrect to suppose that \textit{wh}-misidentification \textit{just is} mispredication; for not just any form of error in predication will count. Returning to Example 8, suppose I say, “That is a large creature,” pointing to something in the distance which in fact is not a creature at all, but an inanimate statue. I judge falsely, but not as a result of any misidentification. I do not misidentify \textit{that} (which I see in the distance) as the \textit{large creature} (if any, nearby). If upon approaching I suspect that it is a statue and not a living thing, I gain evidence that defeats my target demonstrative belief \textit{That is a large creature} but leaves no grounds for an existential remainder \textit{Something is a large creature}. That is, it leaves no part of the \textit{original} grounds for believing the}
One last distinction deserves notice. There are two ways to \(wh\)-misidentify the F. The subject may \(wh\)-misidentify because he mistakenly believes, of the target object \(a\), that it is F, when in fact it is not (Examples 6–8). Alternatively, he may mistakenly believe of \(a\) that it is \(not\) F, when in fact it is. This too counts as \(wh\)-misidentification, since it involves failure to figure out which thing is F. To illustrate:

*Example 9.* I am shown a photograph of some colorful geysers and told, “This is America’s first national park.” Having never been there myself, I speculate over its geographical location. Not seeing any mountains, I come to believe (falsely, though justifiably) of America’s first national park, wherever it is, that it is not within the Rocky Mountains. Unbeknownst to me, it is.

*Example 10.* Inspired by the photograph, I decide to take a trip out to Yellowstone National Park. I come to falsely but justifiably believe of this place (Yellowstone) that it is not within the Rockies. My friend, amused at my geographical ignorance, informs me that we *are* in the Rockies.

In both instances, although my target belief is of the form ‘a is not F’ (I mistakenly believe of a certain place that they are *not* a part of the Rockies), I still make an error of \(wh\)-misidentification by failing to figure out where are the Rocky Mountains. Yet in both cases no identity assumptions are in play. My judgments are not based on reidentifying anything as *America’s first national park* or as *this place*. Their difference however lies in whether a *de re* belief is involved. In Example 9, I am unable to hold any *de re* beliefs about America’s first national park, having never seen it myself. In Example 10, I am physically there; so *de re* belief is possible after all. In each case, I am still entitled to a remainder belief to the effect existential. I may have other grounds for believing that something else in the vicinity is a large creature (perhaps right behind it!). I may even have other grounds for believing *That is a large creature* despite my own doubts (as when you say you think you saw it move). But these other independent grounds in no way make my judgment, as formed solely on the basis of what I see and believe about it, liable to error through \(wh\)-misidentification.

---

\(^{21}\) Examples 9 and 10 below demonstrate how this distinction concerning the form of the target belief (‘a is F’ vs. ‘a is not F’) cuts across our earlier distinction concerning the *grounds* for belief (enabling *de re* belief or not).

\(^{22}\) Not even any assumptions of *non*-identity. See §4.

\(^{23}\) Some may insist a photographic picture really *does* enable *de re* thought about the thing represented. We can easily alter the example: “Suppose I merely heard rumors about America’s first national park…”

---

12
that someplace is within the Rockies, even after being corrected of my geographical ignorance. In fact, the defeating evidence in each instance would rebut my original target belief, giving me grounds for believing the opposite about the place in question (of America’s first national park or of this place, that it is within the Rockies).

*Wh*-misidentification is also possible for negative attributions made in the first person.24

*Example 11.* After a long day’s hike, sitting in the car, I catch a whiff of body odor. “Who forgot to wear deodorant?” I ask. In interrogating others, I presuppose that whoever it is, it cannot be me who smells so bad. Upon closer examination, I realize, to my horror, that in fact I am the one who smells so bad.

At first I deny being the smelly offender. My denial is mistaken. I have *wh*-misidentified myself as not being the one who smells. My grounds for denial are lost when I discover the truth. Yet defeating evidence leaves grounds for a ‘remainder’ belief that someone smells bad—namely, myself.25 As before, there are two ways to *wh*-misidentify in the first person: I may mistakenly take myself to not be the F when in fact I am (as in Example 11) or I may mistakenly take myself to be the F when in fact I am not (suppose, alternatively, that I suspect I am the one who smells bad when really it is someone else). Yet unlike before, when it comes to first-personal *wh*-misidentification, the *de re* belief is not optional: one must obviously be in a position to make *de re* judgments about one’s self in order to mistake himself as either being or not being the F.

Examples 9–11 show that *wh*-misidentification is possible even for judgments of the form ‘a is not F’. This fact goes unaccounted for by Pryor’s definitions. In contrast, the account of *wh*-misidentification offered here correctly handles cases like these (via the parenthetical clauses concerning “is not F”). The existence of such cases is nontrivial. First, *wh*-misidentified judgments of this form do not behave

---

24 This is an altered version of Perry’s (1979) classic sugar trail example.

25 “…‘remainder’…” (in quotes), because these are strictly speaking not part of my original grounds, but instead part of my newly gained defeating evidence.
exactly like ordinary cases of \textit{wh}-misidentification. One point of difference is in their remainder belief, which involves predicing Fness rather than non-Fness. This shall become important later on. Second, many introspective present-tense psychological \textit{negative} self-ascriptions (of the form ‘I am \textit{not} F’) have as much a claim to \textit{wh}-IEM as do their ordinary positive correlatives (‘I am F’). This spans a wide range of first-personal judgments, including disavowals of sensations or feelings (“I am not in pain”, “I don’t feel anything in my foot”), moods or emotions (“I’m terribly unhappy”, “I’m not worried about tomorrow”), intentions (“I have no intention of continuing our conversation”), desires or likings (“I don’t particularly enjoy the sensation of free fall”), and beliefs (“I’m agnostic on that issue”). Neglecting these sorts of judgments, as we shall see, risks overlooking counterexamples to Immunity.

III. Thought insertion

In cases of thought insertion, patients report experiencing certain thoughts they claim are not theirs (see Frith (1992)). Campbell suggests they commit an error misidentification:

“What is so striking about the phenomenon of thought insertion as described by schizophrenic patients is that it seems to involve an error of identification. The patient might say, “Thoughts come into my head like ‘Kill God’. It’s just like my mind working, but it isn’t. They come from this chap, Chris. They’re his thoughts.” [Frith 1992: 66] A patient who supposes that thoughts have been inserted into his mind by someone else is right about which thoughts they are, but wrong about whose thoughts they are.” (Campbell 1999a: 609–10)

What is so unusual about inserted thoughts is how the subject denies ownership of them and instead attributes them to someone else, despite acknowledging that these thoughts and their contents are directly available to introspection. Of course, the patient never denies being the \textit{thinker} of his thoughts, in the sense of \textit{experiencing} them. Even inserted thoughts get tokened, as it were, in the mind of the patient. Rather, the schizophrenic patient denies ownership of the thought in some stronger sense.
Following Campbell (2002), let us distinguish two possible senses in which one may be the thinker of a thought. One is as author; the other is as recipient. To be the author of a thought is to be the producer of that thought. To be the recipient of a thought is to entertain or somehow be directly aware of it. It is not enough to secure authorship that a thought merely be tokened by some subject: one has to be the actual source of its production. A subject may conceivably take himself to be a mere recipient of a token thought: he finds himself directly aware of it and able to think it, introspect its contents, and so on; he just doesn’t take himself to have produced it. In that case, he takes himself to be its recipient but not its author.

This is exactly the situation of the schizophrenic subject. Whether or not it is actually possible for authorship to be divorced from recipiency in this way is a further issue. It seems clear enough though that the patient takes himself to be in such a situation. He means to deny being the author of his inserted thoughts. He does not deny being the recipient of those thoughts. (Indeed, he laments being the recipient—and quite an unwilling one!) This respects the patient’s self-reports. The schizophrenic is not unaware of the oddity of his own situation: of having direct access to what seems to be another’s mental state, in much the same way one is acquainted with one’s own.

The alleged status of IEM at stake for our conception of the first person concerns the first-personal notion of thinker in this stronger sense of authorship. I shall argue that thought insertion poses a putative counterexample to the thesis of Immunity that all judgments about one’s own thoughts and other psychological states made on the basis of introspection exhibit IEM. Having differentiated wh-IEM from de re IEM, we may disambiguate between two readings of that claim:

---

26 The thought is already anticipated by Campbell (1999a:610) when, in a similar vein, he distinguishes two senses of owning a thought: the inserted thought is his insofar as he has “some especially direct knowledge of it”, although “there is a sense in which the thought, as it were, remains the property of someone else.”

27 So understood, in attributing the thought T to Chris rather than himself, the subject really takes the author of T to be Chris rather than himself. Whether he also takes Chris to be a recipient of T is a further empirical question.
Any introspection-based belief about one’s own occur
tent psychological states is immune to error through de re misidentification.

Any introspection-based belief about one’s own occur
tent psychological states is immune to error through wh-misidentification.

It is important to distinguish these two readings. Although nontrivial, their difference goes largely unacknowledged in the Immunity literature, which to date has ignored the distinction between de re IEM and wh-IEM. Both parties in the Campbell-Coliva debate have failed to identify exactly which claim is under dispute, de re Immunity or wh-Immunity. Insofar as attempts at further precision are made, these tend to make de re IEM the target notion and hence de re Immunity the target claim. This tendency is unjustified. It is rather, as we shall see, the weaker claim of wh-Immunity that is under challenge from the phenomenon of thought insertion.

Before continuing, I wish to draw attention to another common, yet confusing, pattern found in the Immunity debate. This is the tendency to appeal to a certain form of criterion used by Evans (1982) and Shoemaker (1969) in determining whether or not a given judgment shows IEM. According to the Evans-Shoemaker test, my first-personal judgment that ‘I am F’ exhibits IEM just in case, assuming information is being gained in the normal way, I cannot sensibly ask, “Something is F, but am I F?”28 Two concerns are worth raising. First, this by no means offers any substantive definition or account of IEM, but only a rough-and-ready test for when IEM is present. Given its reliance on the caveat about “normal” conditions, there is no guarantee of its reliability in “abnormal” situations like that of thought insertion. In contrast, the account presented here for both types of misidentification (de re and wh-) makes no

---

28 Shoemaker (1970) used this test to conclude that, given the possibility of quasi-memories, our ordinary apparent memories are in fact not immune to error through misidentification. His argument essentially says that the scope of Immunity should exclude first-person memory-based judgments. Evans disagreed (see 1982: §§6.6, 7.2, 7.3, 7.5). Claims like Immunity, for Evans, are properly construed as generic claims restricted to “normal circumstances” of judgment making. As such, they do not concern “abnormal” belief-forming processes such as those that produce quasi-memories. Evans accordingly revised the ‘coherence’ test for IEM by adding the assumption of normality. Both Shoemaker and Evans believed such a test is available for IEM and went on to deploy it in the debate over memory IEM, albeit drawing different conclusions; their methodology, however, was the same. Yet it is this shared methodology which I find suspect (see below).
reference to the murky notion of normality. Second, as Pryor (1999: §IV) points out, the Evans-Shoemaker test does not even test for de re IEM; what it tests for, rather, is wh-IEM. Combined, these two tendencies compound confusion: it appears debaters over Immunity have been employing the wrong sort of test for the sort of target claim they have in mind—that is, de re IEM (see n.3).

Let us put aside de re Immunity for now (we shall return to it later). I shall argue that thought insertion refutes wh-Immunity: the schizophrenic makes an error of wh-misidentification when reporting on his own inserted thoughts. In disavowing the thought T by saying it was inserted into his mind and so attributing it to someone else (here, Chris), the subject maintains both:

(1) \( T \) is Chris’s thought

(2) \( T \) is not my thought

The subject reports (1) and (2) on the basis of introspection—again, where these are to be understood as claims of authorship, not recipiency. He ascribes the inserted thought T to someone else, despite having direct knowledge of its contents, in much the same way he has direct epistemic access to the contents of thoughts that are really his own (and acknowledged as such). This is symptomatic of his disease. Our ordinary conception of introspection does not countenance the possibility of directly introspecting another’s mental states or of disavowing what states are offered to direct introspection. Yet the subject takes himself to have introspective grounds for believing (1) and (2).

I propose that the subject’s introspection-based beliefs in (1) and (2) are both false on account of wh-misidentification. Neither belief shows wh-IEM.

---

29 Campbell at times appears to have in mind the notion of wh-misidentification, as when he says that the subjects correctly report what thoughts these are, but fail to figure out “whose thoughts they are” (1999a: 610).

30 “Directly introspecting”, since introspection does arguably at times offer knowledge of another’s mental states, e.g. when I recall, through introspection, memories of what other people believed.

31 Note: it is not the inserted thoughts themselves which I claim to be vulnerable to wh-misidentification, but rather beliefs made on the basis of introspection about who is thinking them.
Consider his belief in (1). This belief is untrue, because it falsely presupposes that T belongs to someone named Chris other than the subject himself, when there is no such individual. (I am assuming that “Chris” in (1) is an empty proper name, though nothing essential rests on this assumption.\textsuperscript{32}) At best the subject has defeasible grounds for believing (1). Suppose the subject is informed that Chris does not really exist, but is an imaginary byproduct of his unfortunate pathology; that those in his condition are simply susceptible to such mistaken, however justified, beliefs. The subject may or may not believe this explanation, depending on his degree of delusion or irrationality (an empirical question I leave aside). But supposing he does, this serves as an undercutting defeater to his belief in (1): the subject loses his grounds for believing that T belongs to Chris. Yet arguably he still has grounds for saying nonetheless that these thoughts must belong to somebody. That is, he remains justified in believing:

(3) T is somebody’s thought

The grounds for his target belief, although defeated, leave an existential remainder. It is not unreasonable to think the subject can continue to make the existential claim based on introspection.\textsuperscript{33} For the tokening of the thought is indisputable. After all, the schizophrenic patient is perfectly able to make out what its contents are (“Kill God”).\textsuperscript{34} And every thought originates with someone—at least, this is presupposed by the subject’s disavowing his inserted thoughts while simultaneously attributing them to others. Moreover, there is no principled reason to rule out his reports as contradictory or incoherent (see Campbell 1999a: 610–11; also §§4–5).

\textsuperscript{32} On some views, since “Chris” fails to refer to anything, (the belief that) (1) is false, hence untrue. On other views, the false presupposition renders the entire statement (belief) indeterminate, hence untrue. Either way, it is untrue. The assumption that “Chris” in (1) functions as a directly referential term is inessential. Even if it is a disguised description or descriptive name, going proxy for something like whoever else this thought belongs to, since nothing answers to that descriptive content, (the belief that) (1) is again false, hence untrue.

\textsuperscript{33} Even our subject, we may suppose, refuses to countenance Lichtenberg’s possibility of there being thinking or some token thought without a thinker, despite admitting the possibility of thinking another person’s thoughts.

\textsuperscript{34} The exact content of the thought does not matter. Nor must we even assume it is an inserted thought. Other psychological states may be disavowed and misattributed—e.g. the feeling of an emotion (intense anger) or the performance of an action, whether intentional (talking) or not (crying).
Clearly, the subject’s grounds for the target belief in (1) rest on his grounds for the existential belief in (3). Any undercutting defeater to the latter defeats the former. For the belief that \( T \) is Chris’ thought presupposes the existential assumption that \( T \) is somebody’s thought. Suppose our subject suffers not just from schizophrenia but also skepticism: dubious of any claims to Cartesian incorrigibility, he questions the very occurrence of his inserted thoughts, and no longer takes himself to be justified in believing (3) on introspective grounds. This would undercut his introspective grounds for believing (1) too.

The converse order of dependence fails. The subject’s introspective grounds for believing (1) do not wholly rest upon, but go beyond, his introspective grounds for believing (3).\(^{35}\) The former may be undercut while leaving the latter intact, as when the schizophrenic is informed about his own condition.

Similarly for the subject’s belief in (2). The subject denies the thought is his. Whoever’s it is, he figures, it is someone else, not himself. That much, he claims, is delivered by introspection. Yet these introspective grounds for believing (2) are defeasible. The subject who becomes self-aware of his own pathological state may no longer take himself to be justified in disclaiming these thoughts. This would undercut a portion, but not all, of his justification: his remaining grounds would still justify him in holding onto the existential content of his belief: that \( T \) belongs to somebody. Thus his introspective grounds for believing (2) include, but exceed, his original introspective grounds for believing (3): the former may be undercut without undercutting the latter, but not vice versa. The grounds he possesses for believing the bare existential claim about \( T \) belonging to somebody leave it open as to whether or not he is the owner; taken alone, they provide him no grounds for either believing or disbelieving (2).\(^{36}\)

Both beliefs, in (1) and (2), fail to satisfy the conditions for \( \text{wh-IEM} \). Thus the subject may be said to \( \text{wh-misidentify} \) the thinker of his inserted thoughts when he disowns them and ascribes them to others.

\(^{35}\) “…introspective grounds”, since learning, by way of observation or testimony, of the presence of someone else (perhaps also named Chris) thinking \( T \) or some thought with the same content might offer him other, non-introspective grounds for believing what he does.

\(^{36}\) Additional grounds for believing that no one else is the one thinking \( T \) are required in order to justify him in believing the stronger claim that he is the one thinking \( T \) after all. Without any such further justification, the subject may be left in a state of agnosticism: certain that \( \text{someone} \) is thinking certain thoughts, but uncertain about \( \text{who} \) it is.
IV. Negative attributions

To recap: I have argued that inserted thought reports exhibit wh-misidentification. The subject fails to figure out who is the author of such thoughts when he disavows them and attributes them to others.

Interestingly, Pryor’s account of misidentification fails to deliver these results. A brief word on why: According to Pryor’s original definition (see n.18), when one wh-misidentifies some object a as the F, undercutting one’s mistaken belief that a is F leaves behind an existential remainder that something is F. But the schizophrenic subject’s belief in (2), to the effect that he is not the one thinking the thought T (in the authorship sense), is of the form ‘a is not F’, and not ‘a is F’. Substituting ‘not F’ for ‘F’ in Pryor’s definition, however, results in a remainder belief of the form ‘something is not F’. This predicts, when asked what grounds for belief are left, that the subject (assuming he wh-misidentifies) still has grounds for believing that somebody is not thinking T. And that is the wrong answer.37

In contrast, my proposed formulation allows the target belief to have either form: ‘a is F’ or ‘a is not F’. Both cases share in common an existential remainder of the form ‘something is F’. This delivers the right results for both the schizophrenia cases and ordinary cases of wh-misidentification (Examples 6–8). Hence the case I am making against wh-Immunity crucially relies on my proposed extension of Pryor’s definition of wh-misidentification. This revision is justified, however, given that Pryor’s account fails to account for negative attribution cases of wh-misidentification (Examples 9–11). In fact, it probably better captures what Pryor wanted in the first place: the intended notion of wh-misidentification has it that one’s

37 Similar remarks would apply to de re misidentification (although I do not believe that the error in question is one of de re misidentification; see below). Pryor’s definition of de re misidentification has it that any undercutter to one’s belief that a is F leaves a de re remainder belief that x is F. Substituting ‘not F’ for ‘F’, we get: the remainder for a belief of the form ‘a is not F’ would be ‘x is not F’. But this predicts that, should the schizophrenic subject de re misidentify in believing (3), he would still have grounds for believing that Chris is not thinking T, which again is the wrong answer.
grounds for believing the existential leave it open to him as to whether or not some given object \( a \) is \( F \)—hence in particular, whether \( a \) is not \( F \).

The present argument only challenges \( wh \)-Immunity. It leaves the status of \( de \ re \) Immunity unquestioned. In fact, I claim, somewhat tentatively, that thought insertion poses no counterexample to the alleged \( de \ re \) IEM of introspection-based present-tense psychological self-ascriptions. For the schizophrenic subject does not assume that he himself is Chris. Indeed, no such assumption is possible, given that no \( de \ re \) beliefs can be formed about Chris, who is nonexistent. The absence of any relevant identity assumption or \( de \ re \) belief means this cannot be a case of \( de \ re \) misidentification, as we have defined it.

Nor does it appear to count as \( de \ re \) misidentifying on any suitably amended definition. One will perhaps object: “The relevant assumption here is not one of identity, but of non-identity. The subject bases his belief that \( T \) is not my thought on the assumptions that \( T \) is Chris’ thought and I am not Chris.” To be sure, we can easily extend Pryor’s original account of \( de \ re \) misidentification to cover cases of negative attributions, just as we did for \( wh \)-misidentification, as follows.

**De re misidentification**: One’s grounds for his \( de \ re \) belief about an object \( a \) to the effect that it is \( F \) (is not \( F \)) rests on his grounds for believing, of \( x \) and of \( a \), that \( x \) is \( F \) and that \( x = a \) (\( x \neq a \)), but unbeknownst to him, \( x \neq a \) (\( x = a \)).

As illustrated by a piece of Jewish history:

**Example 12.** King David of ancient Israel conspires to orchestrate the death of one of his commanders in battle, making it appear an accident, in order to take his wife as his own. His prophetic counselor Nathan later comes to him with a moving story about a rich man taking away from a poor man the one lamb he owns and cherishes, only to slaughter it.

---

38 That is because to wonder whether \( p \) just is to wonder whether not-\( p \) (barring intuitionistic cavils about invariance under double negation). “Wondering whether \( p \)” is in this regard unlike e.g. “doubting that \( p \)” or “undercutting evidence for \( p \)”, which are not the same as “doubting that not-\( p \)” or “undercutting evidence for not-\( p \)”, respectively. Yet these latter notions are those used in Pryor’s original characterization of \( wh \)-misidentification (see n.15).
Incited to anger, David says that that man deserves to die. Nathan immediately confronts him. “You are the man!” he exclaims, exposing the king’s own complicity.

David initially expresses his belief that *That man deserves to die*. Before the moment of confrontation, he tacitly believes that *I do not deserve to die*, on the assumption that *I am not that man*. Nathan’s revelation defeats that non-identity assumption. Although David can no longer justifiably believe of *himself* that he does not deserve to die, he is still justified in holding a remainder belief about *that man* to the effect that *he* deserves to die.\(^{39}\) (In fact, in light of Nathan’s rebutting evidence, David must accept the even stronger ‘remainder’ conclusion that he himself deserves to die; or put in first-personal terms: that *I deserve to die*, on grounds of the identity fact that *I am that man*.)

Perhaps, the suggestion then goes, the schizophrenic subject’s denial of his own allegedly inserted thoughts are liable to *de re* misidentification of this modified sort: he denies these thoughts are his on grounds of a mistaken assumption that he is not Chris.

The suggestion remains problematic, for at least three reasons.

First—inserted thought reports still fail to display the sort of *de re* assumptions that are the defining marks of *de re* misidentification. Just as one cannot reidentify a nonexistent individual, it is impossible to discriminate a nonexistent individual. It is not as if the subject is distinguishing himself from, or assuming he is distinct from, some individual Chris he previously encountered, and goes on to attribute the inserted thoughts to *that individual* rather than himself. Chris does not exist. The subject is not in a position to make any *de re* assumptions concerning Chris, identity or non-identity assumptions alike. He ascribes the thoughts to someone else, but there is no one else whom he ascribes those thoughts to.

Second—this result does not depend on the reference failure of ‘Chris’. Even if the schizophrenic ascribes his inserted thoughts to an actual living individual (say, his friend) named Chris, he makes no mistaken identity or non-identity assumptions. He certainly does not mistake himself to be Chris. It is

\(^{39}\) Again, Pryor’s definitions fall short of this result. The *de re* remainder for the belief that *I do not deserve to die* should be *That man deserves to die*. But his account wrongly predicts it to be *That man does not deserve to die* (see n.37).
not as if the schizophrenic patient tries to reidentify himself as Chris (as when I point myself out in an old photographs and remark, “I was a chubby kid”). Nor is it the case that he tries to reidentify Chris as himself (as when the deluded man who takes himself to be Arnold Schwarzenegger makes false judgments—like “Arnold feels powerful”, “Arnold is displeased”, “Arnold wishes for adventure”—all on the basis of true self-ascriptions, i.e. that he himself is all these things). Nor does he make any mistakes in discriminating himself from others (as when a spectator sees a man tarred and feathered and remarks, “I am glad I am not that man”). Indeed, when disavowing his inserted thoughts, he assumes he is not his friend Chris. But he’s right in thinking that—he isn’t Chris. No errors of de re misidentification arise.

Third—perhaps less importantly although nontrivially—the very characterization of de re misidentification* is problematic. Admittedly, it does not follow from \( x \neq a \) and \( x \) is \( F \) that \( a \) is not \( F \). I can justifiably believe that everyone is mortal, including myself; someone else’s mortality is not proof of my own immortality. This marks a difference between judgments based on identity versus non-identity assumptions. For identity assumptions: one’s grounds for believing \( x = a \) and grounds for believing \( x \) is \( F \) together thereby constitute grounds for believing, or offer one knowledge that, \( a \) is \( F \).\(^{40}\) Whereas for non-identity assumptions: grounds for believing \( x \neq a \) and grounds for believing \( x \) is \( F \) do not jointly guarantee grounds for, or otherwise mandate, believing \( a \) is not \( F \)—unless, that is, one has grounds for believing there is at most one \( F \), in which case \( x \) is not just an \( F \) but the only \( F \).

This last uniqueness condition—that one have grounds for assuming there is a unique \( F \)—is required for misidentification, understood in the broadest \( wh \)-sense. The very possibility of \( wh \)-misidentifying, or failing to figure out which thing is the \( F \), presupposes there is a unique \( F \). Yet when is the uniqueness condition guaranteed? It is not guaranteed to be satisfied simply whenever one’s grounds for the target belief \( a \) is not \( F \) happen to rest on his grounds for the de re assumptions \( x \neq a \) and \( x \) is \( F \). Consider:

---

\(^{40}\) This assumes—uncontroversially, for our purposes—some epistemic closure principle (to the effect that warrant is preserved over entailment) and the indiscernability of identicals (by which \( x = a \) and \( x \) is \( F \) jointly entail \( a \) is \( F \)).
Example 13. A stranger nearby is drowning. Selfishly, I reason that so long as there are others there to help, I myself need not bother. I spot two shadows of what I take to be other people. Assuming they will offer help, I decide not to offer any help myself. Now, one of those shadows really does belong to someone else. But the other shadow, unbeknownst to me, is really my own. Luckily, the other person offers assistance (and goes on to rescue the victim). Meanwhile, I stand by silently, offering no help.

My belief that *I will not offer help* is based on the *de re* assumptions, about *x* and *y* whose shadows I see, that *x* will offer help, that *y* will offer help, that I am not *x*, and that I am not *y*, where *x* turns out to be myself. Suppose person *y* does offer to help the victim, while I do not. Now, my target belief does rest on these *de re* assumptions, some of which are false (that *x* ≠ *me* and that *x* will offer help). Yet the uniqueness condition fails: I make no assumptions that there must be at most one rescuer who offers to help. The suggestion about *de re* misidentification* therefore misclassifies Example 13 as an instance of misidentification. To be sure, my ‘intermediate’ belief that *x* will offer help *would* suffer from misidentification of the *de re* sort, since its grounds rest on a mistaken non-identity assumption that I am not *x*. 41 But my target judgment that *I will not offer help*, though it too rests on that same mistaken non-identity assumption, does not itself go wrong on account of misidentification. For in making that judgment, I do not fail to figure out who the rescuer is, since I do not assume there is at most a *single* rescuer. I do not even make a false judgment. For it is *true* that I will not offer help: my selfish intentions prevail. If anything, I correctly identify myself as *not* being the rescuer.

The task of stating exactly when the uniqueness condition is met faces some difficulty. Yet satisfying this condition is required for *wh*-misidentification, and hence also required for *de re* misidentification. 42 The suggested characterization of *de re* misidentification* thus remains an imperfect proposal for how to

---

41 “…*would*…”, because I may have no standing belief that *x* will offer help, only an assumption to that effect (as when my target judgment is arrived at non-inferentially).

42 Recall that any case of *de re* misidentification is a case of *wh*-misidentification. Hence any necessary condition for being a case of *wh*-misidentification is thereby a necessary condition for being a case of *de re* misidentification.
extend the notion of *de re* misidentification to negative attributions: certain necessary conditions remain unstated.\textsuperscript{43} Pending improvement, the intended notion of *de re* misidentification lacks full articulation.\textsuperscript{44}

V. Consequences

In this paper I have so far argued that inserted thought reports demonstrate *wh*-misidentification, not *de re* misidentification. I conclude by drawing several consequences of this observation, concerning the nature of misidentification, IEM and the first person.

*Consequence #1: Reference failure does not explain misidentification*

A semantic account of IEM might attempt to liken ‘Chris’ to other empty terms like ‘Santa Claus’, and explain the failure of IEM in inserted thought reports in terms of failure of reference, as in:

*Example 14.* A little girl wakes up in the middle of the night on Christmas Eve. She sees her father dressed up in a big red suit. The little girl mistakes her father for Santa.\textsuperscript{45} Thinking he is there to put gifts in her stocking, she excitedly exclaims, “Santa’s here!”

\textsuperscript{43} If N is a necessary condition for X and if S\textsubscript{1}…S\textsubscript{n} exhaust all the possible sufficient conditions for N, then S\textsubscript{1}∨…∨S\textsubscript{n} is a necessary condition for X. Therefore any account on which some Si goes unaccounted for will be unable to state a complete set of necessary conditions for X.

\textsuperscript{44} Similar remarks apply for the converse entailment, which also fails: satisfaction of the uniqueness condition does not guarantee that one’s *target grounds* for believing the target claim that *a is not F* rest on his *source grounds* for believing the *de re* assumptions that *x ≠ a* and that *x is F*. Consider:

*Example 13*. As before, I see two shadows belonging to *x* and *y*. Believing the best in others, I assume they will both offer to help the drowning stranger. Knowing my own selfish tendencies, I conclude there is only one coward among us all: myself. Guiltily, I believe that I am the only one around who will not offer to help. It turns out that I am right about *that*, although, unbeknownst to me, I have mistaken *x*’s shadow as someone else’s, when really I am *x*. Person *y* ends up offering to help. I stand by silently.

This time, the uniqueness condition is met, since I believe there will only be one egoist. But my belief that *I (am the only one who) will not help* does not rest on the *de re* assumptions that *I am not x* and that *x will help*. For I may be alerted of the fact that I am *x*, but this would not affect the grounds or truth of my claim to unshared selfishness.

\textsuperscript{45} That is, she falsely believes, of some man, who is in fact her father, that *he* is Santa—not that her father *is* Santa.
The fact that ‘Santa’ lacks a referent explains why she misidentifies the man as Santa: Santa doesn’t exist. One might suspect that the lack of reference for ‘Chris’ similarly explains why its use in inserted thought reports results in misidentification.

This line of semantic explanation is hopeless, for reasons already given. Although the subject’s use of ‘Chris’ fails to refer, wh-misidentification would still arise even if he did successfully intend to refer to some actual individual named Chris. More importantly, ‘I’ as used by the subject does not fail to refer. Reference failure therefore explains neither instance of wh-misidentification: both when attributing one’s inserted thoughts to others and when disavowing those very thoughts. Any reference failure is coincidental and explanatorily inessential—a feature of the disease, not of the referring expression. Unlike with ‘Santa Claus’, it is not as if the schizophrenic is being lied to, manipulated or otherwise misled by other speakers within his community.\(^{46}\) Rather, ‘Chris’ plausibly owes its reference failure to some faulty aspect of the patient’s psychology, such as defective introspective faculties or some other underlying cognitive malfunction (for discussion, see Campbell 1999a, 2002).

Consequence #2: Immunity to reference failure does not explain IEM

Alternatively, some philosophers understand the IEM of first-person judgments as deriving from the impossibility of reference failure for uses of the first-person pronoun (see Peacocke 2012).\(^{47}\) Token uses of ‘I’ can never fail to refer, given that their reference is secured by a token-reflexive rule: any tokening of ‘I’ refers to the speaker, or producer of that token, and no one else. One cannot mistakenly refer to anyone else with his own use of ‘I’. Nor can a tokening of ‘I’ ever fail to refer to anything, assuming that all tokens of ‘I’ are produced. One might think this explains why misidentification is impossible for

\(^{46}\) Empty names like ‘Santa Claus’ presumably enjoy wide public circulation despite being empty for various sociological reasons: parents want their children to behave, tradition dislikes change, kids are easily cheated, etc.

\(^{47}\) Peacocke (2012: 145–8) attempts to give a “metasemantic explanation” of IEM (although see also his 2008: ch.3). Anscombe (1975) also appears to tie together the two purported phenomena: “It seems clear that if “I” is a ‘referring expression’ at all, it has both kinds of guaranteed reference. The object an “I”-user means by it must exist so long as he is using “I”, nor can he take the wrong object to be the object he means by “I”.’ (p.30 from 1981 reprint)
many introspective self-ascriptions: a subject cannot misidentify himself when using a referential device that is solely dedicated to tracking and enabling judgments about a single object—namely, himself. Exceptions occur when reference is unsuccessful. The breakdown of IEM in pathological cases of thought insertion is, on this account, due to problems of mistaken or unsuccessful self-reference.

This semantic explanation for first-personal IEM too is hopeless. We have already seen how inserted thought reports are liable to $wh$-misidentification. Yet there is no reason to believe the subject’s use of the first person is in any way deviant or defective in its reference. His disavowal “T does not belong to me” is truly about himself. It successfully communicates something he believes about his own mental states. Thus, failure of reference does not explain why abnormal self-ascriptions (such as inserted thought reports) are open to misidentification. So there is no reason to think that guarantee of reference should explain why normal introspection-based self-ascriptions are immune to misidentification.

Consequence #3: Immunity to defective reference does not explain IEM

Reference failure is not the only type of defective reference. There is also the phenomenon of missed reference.50

Example 15. Suppose we are at the aquarium. I wish to point out a particularly friendly-looking fish. Turning away from the tank, I point behind me and exclaim, “That is a friendly-looking fish.” Unbeknownst to me, while my back has been turned, the fish I saw was eaten by a bigger, meaner-looking shark.

Whatever mistake has occurred here, it is not one of misidentification, in either the $de$ re sense or the $wh$-sense. It is not as if I attempt to reidentify the friendly-looking fish as the mean-looking shark, or that I

48 Unsurprisingly, the semantic story also fails for other likely IEM candidates: perceptual demonstrative judgments arguably display IEM even though demonstratives are not immune to reference failure. See Campbell (1999b: §III).

49 In fact, Campbell (2004: §3) worries whether the token-reflexive property of ‘I’ actually conflicts with its giving rise to uses that are characteristically IEM.

50 The example is a variant on Pryor’s (1999: 277–8) Spiro-Carnap case of what he calls “badly aimed reference”.
have failed to figure out which creature in the tank is the friendly-looking one. I am simply ignorant of what transpires behind my back. As a result, I accidentally refer to the wrong object. Instead of referring to the fish I have in mind, I end up referring to the shark without knowing it.

One might attempt a similar diagnosis of the error in inserted thought reports. Perhaps when the subject claims, “These thoughts belong to Chris,” he intends to refer to someone else—only, without knowing it, he ends up referring to himself. The case would then be one, not of failed reference with an empty name, but of unintended (self-)reference. (Compare: someone mistakes his own reflection for someone else’s and shouts “Hey you!”) On this account, the subject’s judgments that “T belongs to Chris” and “T does not belong to me” turn out to express inconsistent beliefs about himself.

The suggestion, however coherent, is implausible. Unlike the fish-shark case, it is not as if the subject has swapped places with some nonexistent individual and that this switch has gone unnoticed within his own mind. Rather, there is no one else to refer to in the first place. Why all his uses of ‘Chris’ should somehow be self-referring remains a mystery. Nor can unintended self-reference be the entire story. For recall that inserted thought reports involve two claims: an attribution and a disavowal. An account of the deviant referential features of ‘Chris’, even if successful, still leaves unexplained why the subject misidentifies when denying, in first personal terms, that the thoughts are his.

**Consequence #4: IEM is not a semantic phenomenon**

Semantic treatments of IEM attempt to explain the impossibility of misidentification in normal non-inferential introspection-based psychological ascriptions in terms of special semantic features—i.e. reference-related features—of the terms or concepts in question. Inversely, they attribute the possibility of misidentification in abnormal cases to semantic defects, such as mistaken or failed reference. As the special test case of thought insertion shows, this approach to IEM is highly problematic. The alternative

---

51 The other direction—his use of ‘I’ mistakenly referring to another person named Chris, living or imaginary—is impossible, given the token-reflexive property of ‘I’.
is to construe IEM as a fundamentally epistemic rather than semantic phenomenon: IEM primarily has to do with self-knowledge rather than self-reference. That is the approach endorsed in this paper.

Just as semantic treatments of IEM are found wanting, so are their implications about the first person. Ordinary language philosophers appear to have drawn far-reaching conclusions on the basis of first-personal IEM, such as that uses of the first-personal pronoun are not even referential in many instances (Wittgenstein 1958) or at all (Anscombe 1975). The present considerations suggest it is premature to draw any conclusions concerning the existence or referential availability of the metaphysical self on the basis of semantic considerations alone.

Consequence #5: IEM reveals little about the first person

According to claims like Immunity, a core stock of first-personal psychological ascriptions are guaranteed to exhibit IEM. The significance of this purported fact is disputed. Enthusiasts, following Wittgenstein, believe the alleged IEM reflects some special metaphysical or semantic status for the first person. Deflationists, most prominently Evans (1982), believe it reveals nothing distinctive about the first person.

Wittgenstein (1958: 67) likens the alleged IEM-exhibiting uses of ‘I’ (“as subject”, in his terms) to a sort of conventional groan: “To say, “I have pain” is no more a statement about a particular person than moaning is.” Anscombe (1975) proposes an even more radical noncognitivism about the first person: “I” is neither a name nor another kind of expression whose logical role is to make a reference, at all.” (pp. 30-32 from 1981 reprint) Both were apparently motivated by an aversion to the idea that ‘I’, if indeed referential, would refer to some immaterial Cartesian ego—“something bodiless, which, however, has its seat in our body” (Wittgenstein 1975: 69)—or as Anscombe puts it: “if “I” is a referring expression, then Descartes was right about what the referent was.” (Ibid. p.31) Such views suggest that any alleged first-personal IEM must be illusory, strictly speaking, since the candidate judgments never express truth-evaluable contents—and so say nothing about anyone, not even oneself—at all.

Ordinary language philosophers seemed to treat IEM as an allegedly distinctive feature of the first-personal pronoun (Wittgenstein 1958, Anscombe 1975). Later philosophers, like Shoemaker (1968), recognized that IEM isn’t limited to just first-personal psychological self-ascriptions (but extends to bodily self-ascriptions, demonstrative judgments, etc.), yet nevertheless attributed special semantic status to first-personal reference (although not exclusively through use of the first-personal pronoun; see Shoemaker 1970: fn.5). Peacocke (2012) links this specifically to the token-reflexive reference rule for ‘I’ (although Shoemaker 1968: 560 appears to raise concerns for this sort of approach).
rather, IEM is a general feature of all non-inferential singular thoughts: introspective first-personal thoughts are merely a special case. 55

But as we have seen, thought insertion arguably shows that introspection-based psychological ascriptions are not in general guaranteed to exhibit wh-IEM. This rejection of wh-Immunity favors the deflationist position: introspection is not an essentially privileged mode of knowing, insofar as it too is not impervious to mistakes of wh-misidentification. Inserted thought reports are proof that one can fail to figure out who is the producer of certain mental states delivered by introspection. Yet something of the enthusiast’s hopes remain. The present arguments leave the alleged de re IEM status of introspective psychological self-ascriptions untouched. Perhaps the truth of de re Immunity reflects how introspection never grounds false identity (or non-identity) judgments. My introspective faculties cannot tell me that I am not myself, or that I am someone else. Even so, misidentification is nonetheless possible, just not as a result of a mistaken identity (or non-identity) assumption. It remains up to enthusiasts to show why this should matter in our philosophical investigations of the first person.

55 See Evans (1982: §6.6) and Wright (2012: §3). Although in earlier writing, Wright (1998:§2) appears to view the IEM of other judgments as parasitic on the IEM of a core stock of first-personal psychological self-ascriptions.
References


