FIRST-PERSON AWARENESS OF INTENTIONS AND IMMUNITY TO ERROR THROUGH MISIDENTIFICATION

[PENULTIMATE DRAFT – PLEASE REFER TO PUBLISHED VERSION FOR CITATION PURPOSES]

No-one thinks these days that the mind is transparent to itself. It is generally recognized that parts of one's mental life are hidden from one. Nevertheless, it seems that each of us enjoys a special awareness of at least some of her own mental goings-on that she cannot have of anything or anyone else. Let us call this 'first-person awareness' of one's psychological states. Its nature is disputed. One question concerns its structure. On the dominant view, it is claimed to have an act-object structure, i.e., it is claimed to consist in acts of awareness that have one's mental states as their objects. The relevant form of awareness is non-sensory. It is known as 'introspection'. An alternative view takes first-personal awareness of one's own mental goings-on to have an 'adverbial' structure. To be first-personally aware of a mental state on this picture is for that state to be conscious. But 'conscious' does not denote an act of awareness that has the state as its object. Instead, it describes the sort of state it is. It is possible that first-person awareness is not a unified phenomenon. Its structure may differ for different mental states. Those writers who endorse an adverbial account – e.g., Merleau-Ponty (1962), Sartre (1993), Moran (2001), O'Brien (2003) – disagree over how much of our conscious life fits this model.

In this paper, I will offer an argument in favor of construing first-person awareness of one's intentions adverbially. My argument will hinge on the importance

of a phenomenon known as 'immunity to error through misidentification' (IEM). An ascription of a state or property to oneself is IEM when one cannot be wrong about *who* is in that state or has that property. A self-ascription made on the basis of first-person awareness of that state is traditionally claimed to be IEM. As I will explain below, IEM self-ascriptions of intention are essential to first-person thinking. Since many humans undoubtedly have first-person thoughts, they must be capable of making IEM self-ascriptions of intention. The only possible basis for doing so is first-person awareness of them. Thus I take it that an account of what it is to be first-personally aware of one's intentions should construe it in such a way that it can ground IEM self-ascriptions. I will then consider a challenge posed by Jeannerod and Pacherie (2004) who claim that certain empirical data establish that first-person awareness of one's intentions cannot ground IEM self-ascriptions of them. My discussion will show that their argument is unsuccessful. However, there is a second argument in the vicinity that *does* seem to establish this conclusion. I will then argue that it presupposes that first-person awareness of one's intentions has an act-object structure, i.e., that it is introspection. The view that first-person awareness of intention has an adverbial structure is not vulnerable to this attack. This gives us a significant reason to construe such awareness adverbially.

1. First-person awareness of intentions

In this section, I will set out the two rival views of what it is to be first-personally aware of one's intentions.

The majority of theorists hold that to be first-personally aware of one's intentions is to introspect them. Introspection is conceived as having an act-object structure. It consists in acts of awareness that have one's own mental states as their objects. One's mental states are independent from introspective awareness of them. They exist whether or not one is conscious of them. To be aware of a mental state is for it to 'come into view'. On this model, the structure of first-person awareness is analogous to the structure of perception. To perceive an object is to have an experience *of* it. Seeing a table, e.g., involves an act of awareness that is directed at the table. Just as there are different ways to cash out this basic thought in the case of perception, so too there are different accounts of introspection. Descartes (1996), e.g., conceived of it as a sort of infallible inner vision. The mind, for Descartes, is non-physical, thus one's faculty of introspecting it has no physical basis. Armstrong (1963) in contrast, holds that mental states are brain states. He conceives of introspection as the brain's means of scanning its own states. On this picture, introspection is as fallible and likely to malfunction as one's means of sensing the world – sight, hearing, smell, and so forth. Despite the significant differences between these views, both take introspective awareness to have an act-object structure. To be first-personally aware of an intention on this model is for that intention to be the object at which an act of introspective awareness is directed.

The alternative adverbial view holds that to be first-personally aware of an intention is for that intention to be conscious. 'Conscious' should be understood adverbially as describing the kind of state it is, rather than as denoting awareness one

has *of* that state. To have a conscious intention to walk one's dog, e.g., is to consciously intend to walk one's dog. It is to consciously commit oneself to dog-walking. The adverbial account of conscious intention will perhaps be unfamiliar to many readers. It will help to clarify and motivate it if we consider a problem faced by the act-object model, which the adverbial account can overcome. The problem I have in mind is due to Moran (2001). He argues that the construal of conscious intention as consciousness *of* an intention makes the subject a passive spectator of her own decisions. The subject has special access to her own intentions, but her relation to them is the same as her relation to the intentions of others. The subject 'looks inward' to 'see' what she intends to do in much the same way that she looks 'outward' at Jimmy to see what *he* intends to do. In reality, however, the subject is the *subject* of her conscious intentions – she is the one who *makes* the decisions, who *forms* the intentions – and our conception of conscious intention should capture this fact. Rather than thinking of the subject's awareness as distinct from her intentions, we should think of the awareness and the intention as being part of the same mental phenomenon. The way to do this is to construe 'conscious' adverbially as describing the sort of intention it is: a conscious commitment to a course of action (Moran 2001: 31).

The adverbial account of conscious intention has the following consequence. If one has a course of action 'before one's mind' – i.e., if one consciously entertains it – that course of action will only be a conscious intention if one consciously commits oneself to that course of action at the time that one considers it. This is so, even if the

course of action the subject consciously entertains is one to which she previously committed herself. Moran puts the matter like this. Consciously considering a course of action puts the subject in the position of being able to choose to commit herself to it. What she chooses to do will depend upon what she finds desirable, sensible, and so on. If the course of action considered is one to which the subject previously committed herself, consciously considering it puts her in a position where she can withdraw her commitment to it. Suppose, e.g., that Anna remembers that she plans to visit Jim on Monday – her previous intention to visit Jim occupies her attention. Entertaining this course of action allows Anna to re-evaluate it. She realizes that she will not have time to visit Jim because she has to go to the dentist, and so revises her plans. The course of action she considers – visiting Jim on Monday – is no longer something she intends. If the subject does not withdraw her commitment to the course of action, then she has effectively reaffirmed her commitment to it. It still seems sensible, desirable, morally worthy, etc. at the time that she consciously considers it. The upshot is that on the adverbial model, the status of a conscious intention as an intention is dependent on the subject's consciously committing herself to the course of action it represents at the time that she entertains it. This will be important later.

2. Immunity to error through misidentification, first-person thought, and intentions

In this section, I will sketch an account of the connection between immunity to error

through misidentification, self-ascriptions of intention, and first-person thought. I have defended this account elsewhere (Romdenh-Romluc, forthcoming), and since space prevents me from offering a full defence of this view here, I will assume for present purposes, that it should be accepted.

Self-ascriptions of intention are judgments of the form 'I intend to φ'. First-person awareness of an intention allows one to self-ascribe it. Self-ascriptions of intention made on this basis are traditionally claimed to be IEM – i.e., one cannot be wrong about *who* has the intention. Tradition also has it that self-ascriptions of intention are immune to a second sort of mistake – I cannot be wrong about *what* I intend to do. But it is the first sort of immunity that interests us here. Not all self-ascriptions of intention are IEM. A judgment *a* is *F* will be *open* to error through misidentification if it is possible for me to know that *someone* or something is *F*, but be mistaken about whether that thing or person is *a*. Thus if I judge, 'I intend to φ', I will make an error of misidentification if *someone* intends to φ, but that person is not me. Suppose, e.g., that my psychiatrist tells me that the source of my chronic tardiness is a suppressed intention to lose my job. On the basis of her testimony, I judge, 'I intend to lose my job'. It is possible for my psychiatrist to mix up my notes with someone else's. If this happens, it will not be *me* who intends to lose her job, but the person whose notes are mixed up with mine. It follows that in this case my self-ascription of intention is *not* IEM. A self-ascription will be IEM if it is based on some way of finding out about my own states and properties that only allows me to find out about myself. Thus a self-ascription of intention will be IEM if it is based on a way

of knowing about my intentions that only provides me with knowledge of my own intentions. Since it seems that I can only be first-personally aware of my own intentions, self-ascriptions of intention based on this form of awareness are traditionally claimed to be IEM. In contrast, I can find out about lots of other people's intentions on the basis of my psychiatrist's testimony. This is why self-ascriptions of intention made on this basis are *not* IEM.

An influential line of thought holds that IEM self-ascriptions are essentially connected to first-person thinking. First-person thoughts are those that an English speaker would typically express in language using 'I'. They involve thinking of oneself in a distinctive way in which one cannot think about anything else. First-person thoughts involve a conception of oneself. But not all conceptions of what is in fact oneself are first-personal. Perry's (1979) famous example of the messy shopper illustrates this point. One day whilst out shopping, I notice a trail of sugar on the grocery floor. I think to myself, 'That person with the torn sugar bag is making a mess'. I follow the trail of sugar, intending to tell the messy shopper that her sugar bag is torn. At some point, I realize that the person making the mess is *me*. I think to myself, 'I am making a mess', and stop to adjust the sugar bag in my shopping trolley. In this example, both of my thoughts involve a conception of someone who is in fact myself. But only my second thought is first-personal. It follows that first-person thoughts involve a special kind of self-conception. First-person thoughts are not distinctive because of their object (oneself) – as we have just seen, I can have non-first-personal thoughts about the same object. Their distinctiveness is due to the *way*

in which I think about that object. Most, if not all, theories of first-person thought account for this by holding that the self-conception employed in first-person thinking is based on ways of knowing about myself that cannot provide me with knowledge of anyone else. Since these ways of knowing only allow me to find out about myself, self-ascriptions made on their basis will be IEM.

It is generally acknowledged that first-person thoughts involve thinking of oneself as an individual, i.e., as something that can be distinguished from other individuals, and tracked over time.[1] Many theorists – such as Strawson (1959), Evans (1982), and Carruthers (1996) – have argued that one cannot think of a noncorporeal Cartesian self as an individual, because it is impossible to distinguish one Cartesian self from another. Instead, one can only think of oneself as an individual if one thinks of oneself as a bodily being. It follows that the self-conception employed in first-person thinking must be based on ways of knowing about one's body that cannot provide one with knowledge of anyone else. Moreover, since thinking of oneself as an individual that can be distinguished from others and tracked over time requires a conception of oneself as a coherent, unified entity, the relevant ways of knowing about one's own body must provide one with awareness of it as a coherent, unified entity, rather than a fragmented collection of body-parts.

The obvious candidate for this special way of knowing about one's own bodily self is proprioception, which provides awareness of such things as the position of

[1] Although see Anscombe (1981) for an argument against this view.

one's limbs relative to each other, one's active and passive bodily movements, whether or not one's body is in contact with other surfaces, whether one's limbs are hot or cold, whether parts of one's body itch, tickle, hurt, and so on. Proprioception is not a unitary phenomenon. It includes passive forms of proprioceptive awareness (e.g., awareness of one's body being touched; awareness of one's passive movements), and active proprioceptive awareness, i.e., awareness of one's self-generated movements. Tsakiris et al. (2006) present empirical data that suggest it is the latter – proprioceptive experience of one's actions – that yields awareness of one's bodily self as a coherent, unified entity. More passive forms of proprioception only provide a fragmented awareness of one's body. Tsakiris and his colleagues showed this by inducing a proprioceptive illusion. The subject was shown a video image of her right hand as she usually sees it, expect that the image was not aligned with the real position of her hand. The experimenter then either stroked one of the fingers on the participant's right hand, or asked her to move one of them. The video image was turned off, and the participant was asked to proprioceptively identify the position of her fingers. The subject experienced the finger she had lifted, or that the experimenter had stroked as being closer to the seen video image of her hand. Importantly, however, in cases where the participant's finger had been stroked by the experimenter, the proprioceptively experienced location of her other digits did *not* drift towards the video image. But in cases where the participant had lifted the finger herself, the experienced location of her other digits *did* drift towards the location of the video image. Tsakiris et al. interpret these results as showing that the awareness of

one's own body gained via passive forms of proprioception is localized and fragmented, whilst proprioceptive awareness of one's own actions presents one's body as a coherent, unified whole. They write, 'The active body is experienced as more coherent and unified than the passive body… It seems that the unity of bodily self-consciousness comes from action, and not from sensation' (2006: 431). It follows that proprioceptive experience of one's actions is what provides one with awareness of one's bodily self as a coherent and unified individual entity. It is thus essential for first-person thought. On the usual conception, one's actions are those bodily movements one intends to perform. Awareness of them essentially includes awareness of the intentions that brings them about.

The upshot is as follows. It is clear that we have first-person thoughts. This means that each of us must have some means of finding out about her own states and properties that cannot provide her with knowledge of anyone else. These ways of knowing essentially include the experience of one's willed bodily movements, which involves consciousness of one's own intentions. Thus each of us must have some way of finding out about her own intentions that does not provide her with knowledge of what anyone else intends. A self-ascription will be IEM if it is based on some way of finding out about one's states and properties that only provides one with knowledge of oneself. Therefore, the fact that each of us can engage in first-person thinking means that we must have some means of making IEM self-ascriptions of intention. The only possible basis for doing so is first-personal awareness of one's own intentions. It follows that first-person awareness of one's own intentions must be

capable of grounding IEM self-ascriptions of them.

3. The 'naked intention' argument

Jeannerod and Pacherie (2004) have recently challenged the claim that being first-personally aware of an intention allows one to make an IEM self-ascription of it. The basis for their challenge is the hypothesis that subpersonal simulation underlies our capacity to understand another's intentions. They argue that if this hypothesis is correct, then self-ascriptions of intention made on the basis of first-person awareness are not IEM.

The simulation hypothesis holds that the same brain systems are responsible for both producing my actions, and providing me with a grasp of what others intend when I observe their behavior. Both processes produce representations of action. When I prepare to act myself, representations of action are produced to initiate and control my movements. When I watch another's behavior, the action mechanism in my brain simulates her action, thus generating representations of the sort that initiate and control it. Let us call these representational states 'proto-intentions'.[2] Importantly, a proto-intention with the same content is produced whether I prepare to φ myself, or watch another φ-ing. Further processing is required to either produce an intention that will bring about my action, or provide me with understanding of the

[2] Jeannerod and Pacherie simply call these states 'intentions'. The different terminology I have adopted here makes no difference to their argument.

other's intentions. Various theorists have offered evidence in support of the simulation hypothesis. Experiments conducted on monkeys establish that the same 'mirror neurons' fire whether the animal is performing a particular action itself, or watching another animal perform the same action (Di Pellegrino et al. 1992; Rizzolatti et al. 1995). Mirror neurons are also found in humans (Decety et al. 1994, 1997; Grafton et al. 1996; Rizzolatti et al. 1996; Gérardin et al. 2000; Ruby and Decety 2001; Mukamel et al. 2010).

Jeannerod and Pacherie (2004) argue that the mechanisms responsible for processing the proto-intentions generated by the mirror system can break down. When this happens, they claim that the subject will misattribute intentions on the basis of first-person awareness of them. They take certain pathological cases to be cases where this has happened, and as such they take them to both provide support for the simulation hypothesis, and to be actual counterexamples to the claim that first-person awareness of intentions gives rise to IEM self-ascriptions of them. The cases they have in mind are schizophrenic subjects with 'first-rank symptoms', which involve delusions of being controlled by external forces, or of controlling other people. Jeannerod and Pacherie claim that delusions of the former sort involve misattribution of one's intentions to others, whilst delusions of the latter kind involve misattributions of others' intentions to oneself (2004: 133). They argue further that since it is always possible for these processing mechanisms to fail, *no* self-ascription of intention made on the basis of first-person awareness is IEM. I will return to the pathological evidence a little later. For now, I want to examine their argument in more detail.

The simulation hypothesis holds that proto-intentions with the same content are produced by the mirror system whether one prepares to φ oneself, or watches another φ-ing. Since a proto-intention with the same content is produced in both cases, it seems that proto-intentions represent types of actions, and not the subject who intends to act. In this sense, proto-intentions are 'naked'. The downstream processing mechanism that allegedly breaks down in cases of schizophrenia is responsible for adding a subject to the content of the naked proto-intention. Thus a proto-intention arrives from the mirror system with the content 'x φ's'. The downstream processing mechanism then adds 'I' or 'that person I see' to the naked proto-intention to either yield an intention with the content 'I φ', or a state with the content 'that person I see φ's'. Note that this is not the same as ascribing an intention to someone. An ascription of an intention is a *judgment* that someone intends to φ. It is a second-order state about an intention. The downstream processing mechanism does not produce judgments about the proto-intentions generated by the mirror system. It alters the content of those states. Jeannerod and Pacherie suppose that when the latter mechanism malfunctions, the wrong subject is added to the proto-intention. If the person with the malfunctioning system then becomes first-personally aware of the resulting state, they will attribute the relevant intention to the subject represented by its content. Thus if the person becomes aware of an intention with the content, 'I φ', she will ascribe it to herself, whilst if she becomes aware of a state with the content 'that person I see φ's', she will attribute an intention to φ to that person. But since the processing mechanism has malfunctioned and added the wrong subject

to the state produced by the mirror system, the subject makes a mistake of misidentification when she attributes the intention to φ.

An initial problem with Jeannerod and Pacherie's argument is that they need to provide some account of what it is for the downstream mechanism to add the *correct* subject to a proto-intention. What determines whether the final content should represent *me* φ-ing, or the person I see? One might initially suppose that 'that person I see' should be added to the content in cases where the state has been generated as a result of watching someone else act. But there are cases where this is intuitively false. Suppose, e.g., that I see someone enter my local library. Borrowing books from the local library strikes me as a good idea and so I follow them into the library. The states that guide and control my action of entering the library result from watching another person enter the library. However, it seems obvious that *I* intend to enter the library, and so it seems that the states produced by the mirror system that guide and control my action should be processed to represent *me* as their subject. It may well be that this difficulty is not insurmountable. But I will not discuss the matter further as there is a far more significant worry.

Jeannerod and Pacherie focus on 'intentions-in-action' (2004: 138). These are intentions that bring about action *now* and guide its performance. Jeannerod and Pacherie identify them with motor commands. The problem is that the subject cannot be immediately aware of intentions-in-action in the way that Jeannerod and Pacherie assume. Consider first, what it is like to understand another's intentions by observing his behavior. One perceives his intentions as *manifest* in his behavior. When I see

someone open a window in a stuffy room, e.g., I immediately see him as intending to let in some air; I do not need to *infer* that this is his intention. In fact, Jeannerod and Pacherie accept that this is so. They claim that one *perceives* other people's behavior as intentional, and – where appropriate – as directed towards a goal that one can identify (2004: 138). Given that I understand what another intends by seeing his intentions as manifest in his behavior, it seems that we should understand the simulation hypothesis as claiming that simulation underlies my perception of another's behavior as intentional. When I watch another act, simulation mechanisms generate the states that initiate and guide such behavior, which are then *processed subpersonally* to provide me with a perception of the other's behavior as intentional. On this picture, I am not immediately aware of the states generated by the simulation mechanisms; they are not the right kind of state to be available to the sort of first-person awareness I can have of my personal mental states.

Since preparing to act oneself involves the same mechanisms as simulating another's actions, the intentions-in-action produced by the mirror system in the course of acting oneself must be likewise unavailable. It is plausible to suppose that the intentions-in-action produced as a result of preparing to act oneself are processed subpersonally to contribute to one's global awareness of what one wills and how one's body is moving. Empirical evidence also provides a certain amount of support for the claim that intentions-in-action cannot be introspected. Intentions-in-action are very 'finely-grained'. They represent the movements involved in performing some action in great detail. Experimental evidence cited by Jeannerod and Pacherie shows that the

subject's awareness of what she intends to do is, in contrast, fairly 'coarsely-grained' (2004: 120—1). It is true that there are some cases where one's means of sensing some entity does not provide one with very detailed information about it. My bad eyesight, e.g., provides me with the information that the wallpaper is white with a burgundy pattern, but it does not tell me that the pattern is paisley. We might say in this case, that my bad eyesight provides 'coarse-grained' information about the wallpaper's more 'finely-grained' properties. However, it seems *prima facie* that this is not a possibility in the case of first-person awareness of a mental state. It is reasonable to assume that if a psychological state is available to this form of awareness, then *all* of its content is available. If this assumption is correct, then the fact that the subject's awareness of what she intends to do is not as finely-grained as her intentions-in-action provides support for the claim that intentions-in-action are not immediately available to first-person awareness. I will return to this assumption below.

Since intentions-in-action are not the sort of state that is available to first-person awareness, Jeannerod and Pacherie's argument fails to show that self-ascriptions of intention made on this basis are not IEM.

4. The modified naked intention argument

Jeannerod and Pacherie's argument is unsuccessful because it focuses on intentions-in-action, and these are not available to first-person awareness. To show that first-person awareness does not give rise to IEM self-ascriptions of intention, one needs to

focus on the sort of intentions that *are* available to this form of awareness: the personal-level intentions that might figure in one's practical reasoning. Hereafter, I will simply call them 'intentions'. One might suppose that Jeannerod and Pacherie's argument can simply be read as applying to this sort of intention. Thus one might argue that proto-intentions with the same content are produced by the mirror system whether I intend to φ myself, or watch another and so come to see him as intending to φ. These proto-intentions are naked – their content is 'x φ's'. A downstream processing mechanism adds a subject to the naked state, yielding either an intention with the content 'I φ', or a state with the content 'that person φ's'. The downstream mechanism can malfunction and add the wrong subject to a proto-intention state. When I become first-personally aware of an intention, I will attribute it to the subject represented in its content. If the downstream mechanism has failed, I will attribute it to the wrong subject. Thus I will make an error of misidentification. We saw above that Jeannerod and Pacherie need to provide some account of what it is for the downstream mechanism to add the *correct* subject to a proto-intention. The same is true for the version of the argument we are currently considering. I suggested above that it was not immediately clear how to do this. However, now that we are dealing with personal-level intentions, there is an obvious solution. One's personal-level intentions flow from one's beliefs, desires, and other mental states. I intend, e.g., to one day read the complete works of Frege because I desire to improve my knowledge of Frege's philosophy, and believe that I can do this by reading his complete works. Thus a proto-intention produced by my mirror system is mine – the downstream

mechanism should add 'I' to the naked proto-intention – when it flows from my other mental states. The upshot of this argument is that no self-ascription of intention made on the basis of first-person awareness is IEM, because it is always possible for the downstream mechanism to fail and add the wrong subject to a proto-intention.

5. First-person awareness and the modified naked intention argument

The above argument poses a serious challenge to the claim that self-ascriptions of intention made on the basis of first-person awareness are IEM. As we saw above, IEM self-ascriptions of intention are essentially connected to first-person thought. A creature that can have first-person thoughts about herself must have some means of making IEM self-ascriptions of intention. The only plausible basis for making IEM self-ascriptions of intention is first-person awareness of them. It follows that there must be a flaw in the modified naked intention argument.

At this point, one might offer the following quick response. Jeannerod and Pacherie (2004) take their argument to show that *no* self-ascription of intention is IEM. However, this conclusion is too strong. All the modified naked intention argument establishes is that for those subjects whose intention-processing mechanisms are faulty – i.e., those people who undergo the relevant sorts of schizophrenic delusion – none of *their* self-ascriptions of intention are IEM. For subjects where this mechanism is functioning correctly, their first-person awareness of their intentions *can* ground IEM self-ascriptions of intention. This 'de facto' immunity to error through misidentification is sufficient to underpin the first-person

thinking of normal (i.e., non-schizophrenic) subjects.[3] But this response will not do. The problem is the implication that subjects who undergo the relevant sorts of schizophrenic delusion are incapable of first-person thought. Recall that the capacity to have first-person thoughts requires that the subject have some way of finding out about her own intentions that does not provide her with knowledge of what anyone else intends. The only plausible candidate is first-person awareness of intention. It follows that if one's first-person awareness of intentions does *not* allow for self-ascriptions that are IEM, then one will be incapable of first-person thinking. But people who suffer from schizophrenia can and do have first-person thoughts. There are many, many first-person accounts of delusion in the literature, given by people with ongoing symptoms of schizophrenia. (See, e.g., the series of first-hand accounts published by *Schizophrenia Bulletin* since 1979). Some of these accounts are from people who undergo the sorts of delusion that Jeannerod and Pacherie take to be evidence for a malfunction in the systems responsible for processing proto-intentions – see, e.g., Bockes (1985). These first-person accounts record the authors' first-person thoughts about their experiences. It follows that the quick response to the modified naked intention argument is insufficient. Some further response is required.

The modified naked intention argument holds that there are two situations in which an intention with the content 'I φ' might be produced. The first is a case where I really intend to φ. Let us call this a real intention. The second is a case where the

[3] Thanks to an anonymous referee for this thought.

downstream processing mechanism has malfunctioned. In this situation, I am watching someone else, and my mirror system has produced a proto-intention 'x φ's', but the downstream mechanism has wrongly added 'I' to it. Let us call this a pseudo-intention. The argument presupposes that (i) a distinction can be drawn between being first-personally aware of a real intention, and being first-personally aware of a pseudo-intention; but (ii) the two states are indistinguishable to the subject – she cannot tell on the basis of first-person awareness whether she is conscious of a real or a pseudo-intention. I will argue that whilst we can make sense of (i) and (ii) on the act-object conception of first-person awareness (introspection), we cannot do so on the adverbial model.

It is clear how the act-object conception allows for both of these claims. To be first-personally aware of an intention on this model is to be conscious *of* it. The intention is the object of one's introspective awareness and exists whether or not one is aware of it. On this picture, to be aware of either real intentions or pseudo-intentions is simply to have the state before one's mind. One is aware of a real intention if the object of one's introspective awareness is a real intention, whilst one introspects a pseudo-intention if a state of this sort has come to one's inner attention. Since both kinds of state have the same content, they will 'look' the same in introspection and so be indistinguishable to the introspecting subject.

In contrast, the adverbial model does not allow us to make sense of (i). The adverbial conception holds that to be first-personally aware of an intention is to have a conscious intention, and this consists in consciously committing oneself to a course of

action. It follows that to be first-personally aware of a pseudo-intention on this model will have to involve seeming to consciously commit oneself to φ-ing without really doing so. To make sense of (i), we must distinguish between *really* consciously committing oneself to a course of action, and only *seeming* to do so. However, this is/seems distinction can only be made if committing oneself to a course of action is conceived as independent from one's consciousness of that commitment – an activity that occurs within one, whether or not one is conscious of it. On this view of conscious commitment, sense can be made of it seeming to the subject that she commits herself to φ-ing when no such commitment takes place. (In much the same way that it can seem to the subject that there is a car in front of her when in fact there is not.) But to view conscious commitment in this way is to revert to the act-object model of conscious intention as consciousness *of* an intention. On the adverbial conception, consciously committing to φ-ing is not consciousness of an activity that happens independently of one's awareness of it. Instead, it is a conscious activity. It follows that no distinction can be drawn between *really* consciously committing to φ-ing, and only *seeming* to consciously commit oneself to φ-ing. If the subject has an experience that can be described as consciously committing to φ-ing, then she really has committed herself to φ-ing.

Notice that the kind of commitment to φ-ing that constitutes an intention to φ is a commitment that only the subject can make. To commit oneself to φ-ing in this way is to be the subject of the intention. It follows that only my own conscious intentions can occupy my attention in this way. Thus if I self-ascribe an intention to φ

on the basis of consciously committing myself to φ-ing, my self-ascription will be IEM, which is the claim we want to defend.

The adverbial model helps vindicate the assumption relied upon at the end of section 3: if a mental state is available to first-person awareness, then all of its content is available; first-person awareness of a mental state cannot provide coarse-grained information about that state's more finely-grained content. It only makes sense to suppose that a mental state's content could be partly undetected, like my bad eyesight fails to fully detect the wallpaper's properties, if first-person awareness of it is thought of as awareness *of* a mental state that exists independently of the subject's awareness of it – i.e., if one conceives of first-person awareness on the act-object model. On the adverbial picture, to be first-personally aware of an intention is to consciously commit oneself to the course of action it represents, not to become aware of an independently existing state. This conception of conscious intention does not allow us to distinguish between the intention and the subject's first-personal awareness of it in the way required to make sense of the idea that the subject could be first-personally aware of only some of the state's properties.

6. Schizophrenia and the self-ascription of intentions

Jeannerod and Pacherie (2004) argue that first-person awareness of intentions cannot ground self-ascriptions of intention that are IEM. The naked intention argument is intended to show that if we accept the simulation hypothesis, then we must also accept that errors of misidentification are possible in principle. Jeannerod and Pacherie then

interpret certain pathological cases as instances where such errors actually occur. Understood in this way, the pathological cases are both evidence in favour of the simulation hypothesis, and actual counterexamples to the claim that self-ascriptions of intention based on first-person awareness are IEM. I have argued that the adverbial model of conscious intention allows us to resist the claim that errors of misidentification are possible in principle. This model should then provide an alternative way to interpret the pathological data so that they do not threaten the claim that self-ascriptions of intention based on first-person awareness are IEM. My final task is to examine the pathological cases and show that this is so.

Jeannerod and Pacherie point to two sorts of case: delusions of being controlled by external forces, and delusions of controlling other people. They take an example of the first type of delusion to be where the patient hears voices. Typically, the voices experienced by the subject comment about her and refer to her in the third person, or give her commands and directions for action. The consensus is that the voices correspond to the subject's own inner speech (Evans et al. 2000: 137). Thus Jeannerod and Pacherie claim that the subject *intends* to say what she hears the voices as saying. The subject should therefore be understood as misattributing her own intentions to another, i.e., she should be understood as misidentifying the subject of her intentions. However, it is not at all clear that this is so.

Jeannerod and Pacherie endorse the simulation hypothesis, which holds that proto-intentions with the same content are produced by the mirror system whether I prepare to φ myself, or watch another φ-ing. A mechanism further downstream is then

responsible for adding a subject to the content of the naked proto-intention – 'I' if I am preparing to φ myself, or 'that person I see' if I am watching another person. Jeannerod and Pacherie claim that if the second mechanism breaks down, the wrong subject will be added to the proto-intention. If the subject then becomes first-personally aware of the resulting state and ascribes an intention to φ on this basis, she will make an error of misidentification. As we saw above, the problem with Jeannerod and Pacherie's naked intention argument is that they tacitly conceive of the simulation mechanism as subpersonal. One cannot be first-personally aware of the states produced by the mirror system. Instead, the states generated when one simulates another's actions, or prepares to act oneself, will contribute to one's perception of the other's behavior as intentional, or to one's global awareness of one's willed bodily movements, respectively. Jeannerod and Pacherie may be right that the delusion of hearing voices results from a malfunction in the downstream processing mechanism, so that the wrong subject is added to a proto-intention. However, since the simulation mechanism they describe is *sub*personal, the subject cannot be *first-personally* aware of the resulting state. It is instead plausible to suppose that the faulty state is processed subpersonally to yield the delusional experience of hearing voices.

An error of misidentification is an error of judgment: it involves mistaking one individual for another in the course of judging that they are one and the same. It follows that for a subject to misattribute her own intention to another, she has to judge, '*x* intends to φ'. An error of misidentification occurs if it is not *x*, but someone else who intends to φ. If no such judgment takes place, then there is no possibility of

error through misidentification. This is not to say that there is *immunity* to error through misidentification either. Again, it is judgments that are IEM. If there is no judgment, then the possibility of either error, or immunity to such error, just cannot arise. Hearing a voice does not constitute judging, '*x* intends to φ'. It follows that the subject's auditory hallucination does not involve an error of misidentification, even if the voice corresponds to the subject's own inner speech, and the delusion arises because the wrong subject has been added to the content of a proto-intention.

Nevertheless, the subject *could* attribute an intention on the basis of her auditory experience. Suppose that the subject – call her Joan – hears a voice commanding her to go and help the king. Joan might judge on this basis, '*a* [the entity she hears speaking to her] intended to say, "Joan, you must help the king"'. Since Joan has attributed an intention, we are now dealing with the right kind of thing to be either immune or open to error through misidentification. Assuming that what Joan hears the voice as saying corresponds to her own inner speech, it is *Joan*, not *a*, who intended to say 'Joan, you must help the king', and so when she judges that *a* intended to say 'Joan, you must help the king', Joan seems to have misidentified the subject of the intention. However, in this case, Joan attributes the intention retrospectively on the basis of her auditory experience. But the claim we want to defend holds that it is self-ascriptions of intention based on *first-person awareness* of the intention that are IEM. Thus Joan's faulty ascription of her own intention to *a* does not threaten the claim that self-ascriptions of intention based on first-person awareness of the intention are IEM.

One might try to argue that the simulation mechanisms generate *personal level* states (this is the line taken by the modified naked intention argument). The delusion of hearing voices come about when the wrong subject is added to a proto-intention so that it has the content '*a* φ's', rather than 'I φ'. But rather than being processed subpersonally to yield the experience of hearing a voice, the subject becomes first-personally aware of the faulty state, and on this basis ascribes an intention '*a* intends to φ'. In making this judgment, she makes an error of misidentification, because it is really she who intends to φ. The auditory hallucination is then caused by the fact that she makes this judgment.[4] However, it's hard to see how hearing a voice could be *caused* by wrongly judging that someone intends to φ. It is much more plausible to suppose that auditory hallucinations are brought about by systems malfunctioning at the subpersonal level. More importantly, even if this account of such delusions is correct, it does not force us to give up the claim that self-ascriptions based on first-person awareness are IEM. This thesis does not claim that I can *always* correctly ascribe my intentions. It only claims that self-ascriptions of intention based on first-person awareness of them are IEM. The adverbial model of conscious intention holds that the relevant form of awareness is the experience of consciously committing to φ-ing. Joan's mistaken ascription is not based on this sort of awareness. Thus her mistake does not threaten the claim that first-person awareness of intentions can ground self-ascriptions of them that are IEM.

[4] Thanks to an anonymous referee for this objection.

Jeannerod and Pacherie claim that in delusions of controlling other people, the subject misattributes others' intentions to herself in a way that counts against the thesis that self-ascriptions of intention based on first-person awareness are IEM. This claim runs into similar difficulties.

There is disagreement over the nature of such delusions. Campbell (2001) suggests that most accounts of delusion can be classed as either empiricist or rationalist. Empiricist accounts take delusions to be broadly rational responses to anomalous experience – see, e.g., Maher (1974, 1988). Thus one might suppose that delusions of controlling other people involve the *experience* of controlling them. Jeannerod and Pacherie take such experience to result from a breakdown in the mirror system. When the subject watches the other φ-ing, a proto-intention with the content, '*x* φ's' is produced. The downstream processing mechanism should add the 'the person I see' to this content, but instead adds the subject 'I'. I argued above that their conception of simulation is subpersonal. On this picture, the faulty state is processed subpersonally to yield the experience of controlling others. If we understand matters in this way, the first problem for Jeannerod and Pacherie is that simply having such experience does not constitute judging that, 'I intend to φ', and unless such a judgment is made, we cannot talk about either the possibility of error through misidentification, or immunity to such error.

The schizophrenic subject may, however, self-ascribe an intention on the basis of her experience, so that there *is* judgment. But there are two problems with supposing that her judgment threatens the traditional thesis. To misidentify the person

who intends to φ, *someone* must intend to φ – the error involves being mistaken about who this is. The difficulty is that there does not appear to be anyone who has the intention the schizophrenic subject attributes to herself. Suppose, e.g., that Joan watches Leo walk into a shop and buy a cheese sandwich, experiences herself as controlling his action, and on this basis, self-ascribes the intention 'I intend to make Leo buy a cheese sandwich'. For Joan to misidentify the subject of the intention, *someone* (other than Joan) would have to intend to make Leo buy a cheese sandwich. But no-one has this intention. Certainly, Leo himself does not intend to make Leo buy a cheese sandwich. The content of his intention must be first-personal, i.e., his intention must be 'I buy a cheese sandwich', which is a different intention altogether.[5] Although there is error here, it is not error of misidentification. The second problem is that even if Leo has the intention that Joan attributes to herself, so that Joan makes an error of misidentification when she judges, 'I intend to make Leo buy a cheese sandwich', she attributes this intention to herself retrospectively on the basis of her experience, *not* on the basis of first-person awareness of the intention. Since the traditional thesis holds that it is only self-attributions of intention made on the basis of first-person awareness that are IEM, Joan's judgment is not a counterexample to the traditional thesis.

One might, as before, take the simulation mechanisms to produce personal level states. On this picture, the wrong subject is added to a proto-intention so that it

[5] See Castaneda (1966, 1967, 1968) and Perry (1979).

has the content, 'I φ', instead of the content, 'the person I see φ's'. The subject then becomes first-personally aware of this faulty state, on this basis judges, 'I intend to φ', which then causes her experience of controlling the other. One might try to argue that in making this judgment, the subject makes an error of misidentification, because it is not she, but the person she watches who intends to φ.[6] But there are several problems with this suggestion. First, it is not clear how simply making a false judgment can cause one to undergo delusionary experience. This problem is heightened by the fact that, on this interpretation, there is a mismatch between the content of the intention the subject self-ascribes, and the delusionary experience. In the example we are considering, Joan experiences herself as controlling Leo's cheese-sandwich-buying. But the faulty state produced by the mirror system has the content '*I* buy a cheese sandwich'. Since the self-ascription is supposed to reflect this state's content, Joan must judge, 'I intend to buy a cheese sandwich'. Whilst the action represented by the faulty state and Joan's self-ascription of intention is the same type as the one Joan sees Leo as performing, the content of both the faulty state and the self-ascription makes no reference to Leo. It is thus doubly mysterious as to how Joan's judgment that *she* intends to buy a cheese sandwich can result in the experience of controlling Leo's cheese-sandwich-buying.

Even if a response to this worry is available, there is a far more pressing difficulty. For Joan's judgment to involve an error of misidentification, we must hold

[6] Again, thanks to an anonymous referee for pointing out this possibility.

that Joan does not really intend to buy a cheese sandwich. Instead, the faulty state produced by her mirror system is only a pseudo-intention. Moreover, her awareness of it presents it to her in exactly the same way that she is presented with her real intentions. But we can make no sense of this possibility on the adverbial model of conscious intention. On this view, to have a conscious intention is to consciously commit oneself to φ-ing. One's consciousness and one's commitment to φ-ing (the intention) cannot come apart. If one has the experience of consciously committing to φ-ing, then one really does intend to φ. The status of a conscious intention *as* an intention is dependent on the subject's consciously committing herself to the course of action it represents at the time she entertains it. When Joan becomes first-personally aware of the faulty state generated by her mirror system, she entertains the prospect of cheese-sandwich-buying. She will only have a conscious intention – i.e., she will only have the sort of experience that can ground an IEM self-ascription – if she consciously commits herself to cheese-sandwich-buying when she entertains this course of action. But if Joan does make this conscious commitment, then she really does intend to buy a cheese sandwich. Of course, there is something odd about the way she has formed this intention, yet is it, nevertheless, a real intention. It follows that in self-ascribing the intention, she does not make any error of misidentification.

I have so far taken delusions of controlling others to involve *experience* of controlling what is beyond one's control. However, the rationalist position takes delusion to be primarily a disturbance in the subject's beliefs that can subsequently affect his experience – see, e.g., Campbell (2001). On this model, the delusions of

controlling other people primarily involve believing that one possesses such superhuman powers of control. We must therefore consider whether or not possession of such a belief could result in self-attributions of intentions that threaten the traditional thesis. There are two ways in which possession of beliefs about one's superhuman abilities could ground self-ascriptions of intentions. First, one's beliefs about one's superhuman powers might lead one to conclude that one is responsible for the observed behavior of others/external events, and so self-ascribe an intention. Joan might, e.g., watch Leo buy a cheese sandwich, believe she has the power to control Leo, and so judge, 'I intended to make Leo buy a cheese sandwich'. As a counterexample to the traditional thesis, this scenario runs into the difficulties outlined above: it is unlikely that anyone has the intention Joan self-ascribes, in which case no error of misidentification occurs. Moreover, the intention is self-ascribed retrospectively and not on the basis of first-person awareness of it, so even if there is error through misidentification, the claim that first-person awareness of intentions can ground IEM self-ascriptions of them is not threatened. Second, one's beliefs about one's superhuman powers might lead one to form unusual intentions, i.e., intentions to influence other people, which one then attributes to oneself. Thus, e.g., Joan might believe that she can make Leo buy a cheese sandwich, form an intention to do so, and judge on the basis of first-person awareness, 'I intend to make Leo buy a cheese sandwich'. Clearly such a case does not threaten the traditional thesis because Joan *does* intend to make Leo buy a cheese sandwich, so when she self-ascribes this intention, no error of misidentification occurs.

The empiricist and rationalist analyses of delusion do not exhaust the accounts offered in the literature. However, our discussion of them makes it plausible to suppose that the problems outlined above will be encountered, no matter how one analyses delusions of controlling other people. It follows that the pathological cases considered by Jeannerod and Pacherie are no threat to the claim that self-ascriptions of intention based on first-person awareness are IEM.

7. Conclusion

The dominant act-object view holds that first-person awareness of an intention consists in an act of introspective awareness that has the intention as its object. On the alternative adverbial model, to be first-personally aware of an intention is to consciously commit oneself to a course of action. 'Consciously' does not denote an act of awareness that has the intention as its object. It describes the way in which one commits oneself. In this paper, I have provided an argument for construing first-person awareness of intentions adverbially. Traditionally, it is claimed that first-person awareness of an intention can ground an IEM self-ascription of it. The capacity to make IEM self-ascriptions of intention is essentially connected to the ability to have first-person thoughts. The only plausible basis for making IEM self-ascriptions of intention is first-person awareness of them. It follows that our model of what it is to be first-personally aware of an intention must preserve the fact that it can ground IEM self-ascriptions of intention.

The simulation hypothesis challenges the claim that first-person awareness of

an intention can ground an IEM self-ascription of it. The same states are produced in cases where I intend to φ, and where I come to see another as intending to φ based on his behavior. These proto-intentions are then processed to add a subject to their content. The mechanisms responsible for processing them can break down. When this happens, it is claimed that the subject will misattribute her own intentions to someone else, and another's intentions to herself on the basis of first-person awareness. This argument presupposes that a distinction can be drawn between being first-personally aware of a real intention, and being first-personally aware of a state which represents *me* φ-ing, but which should have represented someone else as intending to φ. I have shown that this claim can only be sustained if one construes first-person awareness of an intention as having an act-object structure. If it is construed adverbially as conscious commitment to a course of action, no sense can be made of the distinction. It follows that the adverbial account of conscious intention is not vulnerable to the challenge posed by the simulation hypothesis. None of my arguments have been directed against the simulation hypothesis *per se*. It could well be the case that subpersonal simulation underlies the perception of others' behavior as intentional, and the breakdown of the simulation/intention-processing mechanisms is responsible for the peculiar experiences and delusions undergone by certain schizophrenic subjects. But as I have shown, acceptance of the simulation hypothesis does not force us to give up the thesis that self-ascriptions of intention based on first-person awareness of them are IEM. We can preserve this claim if we construe first-person awareness of an intention adverbially. This gives us reason to prefer the

adverbial model, over the act-object view. Thus first-person awareness of an intention should be understood as consciously committing oneself to a course of action.

References

Anscombe, G. E. M. (1981) 'The First Person', in her *Metaphysics and the Philosophy of Mind, Collected Papers, Vol. II*. Oxford: Blackwells.

Armstrong, D. M. (1963) 'Is Introspective Knowledge Incorrigible?', *The Philosophical Review* 72: 417—432.

Bockes, Z. (1985) 'Freedom Means Knowing You Have a Choice', *Schizophrenia Bulletin* 11: 487—489.

Campbell, J. (2001) 'Rationality, Meaning, and the Analysis of Delusion', *Philosophy, Psychiatry, and Psychology* 8: 89—100.

Carruthers, P. (1996) *Language, Thought, and Consciousness: an Essay in Philosophical Psychology*. Cambridge: Cambridge University Press.

Castañeda, H. N. (1966) '"He"': a Study in the Logic of Self-Consciousness', *Ratio* 8: 130—157.

Castañeda, H. N. (1967) 'Indicators and Quasi-Indicators', *American Philosophical Quarterly* 4: 85—100.

Castañeda, H. N. (1968) 'On the Logic of Attributions of Self-Knowledge to Others', *The Journal of Philosophy* 65: 439—456.

Decety, J., Perani, D., Jeannerod, M., Bettinardi, V., Tadary, B., Woods, R., Mazziotta, J. C., and Fazio, F. (1994) 'Mapping Motor Representations with PET', *Nature* 371: 600—602.

Decety, J., Grezes, J., Costes, N., Perani, D., Jeannerod, M., Procyk, E., Grassi, F., and Fazio, F. (1997) 'Brain Activity During Observation of Actions: Influence of Action Content and Subject's Strategy', *Brain* 120: 1763—1777.

Descartes, R. (1996) *Meditations on First Philosophy*. Cambridge: Cambridge University Press.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., and Rizzolatti, G. (1992) 'Understanding Motor Events: a Neurophysiological Study', *Experimental Brain Research* 91: 176—180.

Ellis, A. W. and Young, A. W. (1990) 'Accounting for Delusional Misidentifications', *British Journal of Psychiatry* 157: 239—248.

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

Evans, C. L., McGuire, P. K., David, A. S. (2000) 'Is Auditory Imagery Defective in Patients with Auditory Hallucinations?', *Psychological Medicine* 30: 137—148.

Gérardin, E., Sirigu, A., Lehéricy, S., Poline, J-B., Gaymard, B., Marsault, C., Agid, Y. and Le Bihan, D. (2000) 'Partially Overlapping Neural Networks for Real and Imagined Hand Movements', *Cerebral Cortex* 10: 1093—1104.

Grafton, S. T., Arbib, M. A., Fadiga, L., and Rizzolatti, G. (1996) 'Localization of Grasp Representations in Humans by Positron Emission Tomography 2. Observation Compared with Imagination', *Experimental Brain Research* 112: 103—111.

Jeannerod, M., Pacherie, E. (2004) 'Agency, Simulation and Self-Identification', *Mind and Language* 19:113—46.

Maher, B. A. (1974) 'Delusional Thinking and Perceptual Disorder', *Journal of Individual Psychology* 30: 98—113.

Maher, B. A. (1988) 'Anomalous Experience and Delusional Thinking: the Logic of Explanations', In T. F. Oltmanns and B. A. Maher (eds.) *Delusional Beliefs*. Chichester: John Wiley and Sons.

Merleau-Ponty, M. (1962) *Phenomenology of Perception*. London: Routledge.

Moran, R. (2001) *Authority and Estrangement*. Princeton: Princeton University Press.

Mukamel, R., Ekstrom A. D., Fried, I. (2010) 'Single-Neuron Responses in Humans During Execution and Observation of Actions', *Current Biology* 20: 750—756.

O'Brien, L. (2003) 'On Knowing One's Own Actions', In J. Roessler and N. Eilan (eds.) *Agency and Self-Awareness*. Oxford: Clarendon Press.

Pacherie, E. (2000) 'The Content of Intentions', *Mind and Language* 15: 400—432.

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

Rizzolatti, G., Fadiga, L., Gallese, L., and Fogassi, L. (1995) 'Premotor Cortex and the Recognition of Motor Actions', *Cognitive Brain Research* 3: 131—141.

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., and

Fazio, F. (1996) 'Localization of Grasp Representations in Humans by Positron Emission Tomography. 1. Observation versus Execution', *Experimental Brain Research* 111: 246—252.

Romdenh-Romluc, K. (forthcoming) 'Evans, the Sense of Agency, and Proprioception'.

Ruby, P. and Decety, J. (2001) 'Effect of Subjective Perspective During Simulation of Action: a PET Investigation of Agency', *Nature Neurosciences* 4: 546—550.

Sartre, J. P. (1993) *Being and Nothingness*. Washington: Washington Square Press.

Strawson, P. F. (1959) *Individuals*. London: Methuen.

Tsakiris, M., Prabhu, G., Haggard, P. (2006) 'Having a Body v*ersus* Moving Your Body: How Agency Structures Body-Ownership', *Consciousness and Cognition* 15: 423—432.