Self-notions and top-down distortion

Daniel Morgan

Inquiry 60 (2017) (1-21).

Abstract: John Perry offers an unusually substantive, and initially plausible, account of the conceptual role of first-person thought. This paper critiques Perry's account, particularly in what it says about action explanation, and offers a partial alternative. It also identifies three high-level assumptions about what accounts of conceptual roles should look like that plausibly explain why Perry's account goes off track in the ways that it does – this is the top-down distortion of the title. Identifying and arguing against the three assumptions helps point in the direction of a better account.

Introduction

The distinction between first-person thought and thought that is merely about oneself, as well as being of interest to theorists, is marked in our ordinary practice of attitude ascriptions. 'The amnesiac war hero NN remained deflated, because, although he knew that NN was being celebrated, he didn't know *he* was being celebrated' is a kind of attitude ascription that an ordinary speaker might make. But it is associated with truth-conditions, or at least felicity-conditions, that crucially turn on whether the war hero is thinking a first-person thought – i.e. the thought 'I am being celebrated'.

It does not follow that we, either as ordinary thinkers or as theorists, have firm intellectual control on what the notion of a first-person thought comes to. Many notions that we apply with fluency are ones that we would brutalize if we tried to explain (compare Leibniz and Newton's attempts to define the notion of a limit). First-person thought could be like that. Even if there is a natural intentional kind that we are latching onto, we might be latching onto it mainly *because* the kind itself is a very effective reference magnet, and *despite* wild inaccuracy in our attempts to say what we are talking about. Or we might be latching onto it because we see clearly one its faces, while another, more interesting, face remains neglected or radically misconceived.

In particular, we seem to know something fairly robust about first-person thoughts' *pattern of reference* – they are systematically about the subject. But an account that just mentions that pattern of reference does nothing to explain what the *point* of having first-person thoughts is. One can think about oneself without thinking a first-person thought. If there is nothing more to first-person thoughts than their pattern of reference, they should therefore be dispensable, a mere convenience. That seems to be wrong. In saying why first-person thoughts are not dispensable, it is natural to appeal to the idea of their having a certain *conceptual role*. But, in contrast to 'the pattern of reference of first-person

thought', it isn't just obvious what the referent of 'the conceptual role of first-person thought' is.

The starting-point for this paper is the unusually substantive, and initially plausible, account of the conceptual role of first-person thoughts offered by John Perry. Pithily expressed, Perry's account is as follows:

> Self-notions play this special role: they are the repositories for information gained in normally self-informative ways, and the motivators of types of action whose success normally depends on facts about the agent (1990:17).[1]

Here as elsewhere, Perry offers his account of first-person thought as an account of *self-notions*, which are stipulated to be a kind of concrete particular in one's brain. It is doubtful, however, that this framing is essential. If one were a connectionist, or a dualist, and one doubted the existence of any concrete particulars in the brain that correspond in an interesting way with first-person thoughts, one could still accept Perry's basic account. In what follows, I will mostly follow Perry in speaking of self-notions, but nothing turns on this.

In (I), I offer some further motivation for investigating the conceptual role of first-person thought, and for starting with Perry's account in particular. In (II), I provide a brief and mainly positive assessment of what Perry says about the *input component* of first-person though – i.e. roughly, the relationship between first-person thoughts and ways of gaining information or knowledge. In (III), I provide a more detailed critique of what he says about the *output component* of first-person thought – i.e. roughly, the relationship between first-person thoughts and action – as well as a positive proposal. In (IV), I relate the specific points made in the previous two sections to three high-level assumptions about conceptual roles: I claim that Perry's accepting these assumptions is a plausible explanation of why his account goes wrong in the ways that it does. Part of the interest here is diagnostic in relation to Perry's account. But the larger aim is to point in the direction of a better account of the conceptual role first-person thought, one that avoids top-down distortion.

(I) Why do we need more work on the conceptual role of first-person thought? Why focus on Perry?

I claimed above that it isn't obvious what the conceptual role of first-person though is. That might be contested. Consider the following data points. Thinking "I am being pursued by a bear" incentivizes me to run in a way that thinking "NN is being pursued by a bear" does not. Perception directly supports the thought "I am in front of a large oak tree" in a way that it does not directly support the thought "NN is in front of a large oak tree". Suppose one just says: these, and other similar things, constitute the conceptual role of first-person thought.

---

[1] I will focus on the statement of the view in Perry (1990). Subsequent papers by Perry on the same topic – e.g. (2010), (2011) or (2012) – defend the same basic view, and my critique applies equally to them.

It might be that that is the best we can do. But it is natural to want a theory that is more systematic than this, one that abstracts away from particular *examples* of action explanation or of gaining information, and captures what stable contribution self-notions are making (and, we can add, would make to the cognitive economies of first-person thinkers who have never heard of bears or of oak trees). There is a dialectical consideration that supports this desideratum. Some recent theorists have been skeptical that first-person thoughts have *any* particularly special conceptual role. Often, the position has been some of the claims made about the role of first-person thoughts in action-explanation, for example, are true – i.e. there are some genuine data-points. But, it is argued, those data-points fail to establish anything particularly special about first-person thoughts. Analogous claims could be made about, for example, the role 'Superman'-thoughts in action explanation.[2] It is hard to reply to a challenge like this without a statement of what *general* thing the data-points involving first-person thoughts are supposed to establish. Leaving dialectical considerations aside however, the more fundamental point is one about what kind of account would be most intellectually satisfying – a systematic account would be.

Perry's theory seems to fit this bill of being systematic, as witness the fact that it is stated in a couple of lines and without reference to particular examples. It is also, I think, one of the few such accounts of offer. The opening sentences of John Campbell's 2004 are revealing in this regard:

> We can make a distinction between the conceptual role of the first person and the reference of the first person. By "conceptual role" I mean the use that is made of the term: the kind of procedures we use in verifying judgments using the term and the kind of actions we perform on the basis of judgments involving the term. In "Self Notions," Perry talks about conceptual using the phrase "epistemic/pragmatic relations". He says there are "normally self-informative ways of gaining information" and "normally self-dependent ways of acting". These ways of getting information about the self, and ways of acting on the self, constitute what I am calling the conceptual role of the first person (2004: 206).

Campbell does not attempt to give his own account of the conceptual role of the first-person. Rather, he appeals to, and implicitly endorses, Perry's account. That he does this is natural. His main focus in the paper is on a question about the *relationship* between the conceptual role of first-person thoughts and their pattern of reference – which of these two things, he asks, is fundamental, and explains the other. That seems like a relatively *advanced* question. In particular, it is a question that one should only start trying to answer once one already has a fairly firm grip on what the conceptual role of first-person thought is. For Campbell, that firm grip is provided by Perry's account. Campbell does not say why he mentions Perry's account rather than some rival account -- he does not, for example, give an argument in its favor. I think that Campbell probably mentions Perry's account because there aren't any equally substantive rival

[2] See Cappelen and Dever 2013.

accounts for him to appeal to. But – perhaps surprisingly given the lack of rival accounts, and the fact that Perry's account is stated in apparently clear and simple terms – there are also no extended *critical* discussions of Perry's account in the literature. For all these reasons, I think Perry's account is the natural starting-off point for a discussion of the conceptual role of the first-person.

(II) Self-notions: the input component

Perry's view about the input component of self-notions, already quoted above, is that they are 'the repositories for information gained in normally self-informative ways' (1990:17).

Here is the kind of data point that is intended to support Perry's claim. Suppose I am looking at an oak tree. Vision justifies me in judging 'That oak tree is tall and knobbly'. It also justifies me in making a judgment about myself. It justifies me in judging, for example, 'I am in front of a tall and knobbly oak-tree'. It does not, however, justify me in judging me 'NN is in front of a tall and knobbly oak-tree', unless I know that I am NN.

Intuitively, this has something to do with the fact that the following both hold: (i) vision provides information about how things are around thee subject (vision is normally self-informative, in Perry's phrase) and (ii) the pattern of reference of first-person thought is that they are always about the subject. The fact that (i) and (ii) both hold does something to explain why first-person thoughts based on vision are often true.

The natural way to generalize this data point is to suggest, as Perry does, that normally self-informative ways of gaining information support judgments that involve self-notions – this is how we should presumably understand the 'repository' idea. This also seems to fit introspection and internal bodily awareness, for example. Both of these are normally self-informative and they support first-person judgments.

The only question I want to raise at this point about Perry's account is the following: what justifies the 'the' in Perry's claim? What justifies the claim that self-notions are the *unique* repositories for the ways of gaining information that Perry's theory focuses on?

Perry's uses a fairly abstract property -- 'normal self-informativeness' – to pick out the ways of gaining information he is interested in. When we ask *which* ways of gaining information actually have that property, perception, internal bodily awareness and introspection are what spring to mind. But, on reflection, perception, internal bodily awareness and introspection seem "current-moment" informative just as much as they self-informative. Granted, vision doesn't just let me know that *someone* is in front of a tree, it lets me know that *I* am in front of a tree. But, similarly, vision doesn't just let me know that I am, have been, or will be in front of a tree. It lets me know that I am *now* in front of a tree. So, on the face of it, one's now-notions will *also* be repositories for normally self-informative ways of gaining information. If so, then neither self-notions nor now-notions can accurately be described as 'the' repositories for normally self-

informative ways of gaining information. I will return to this challenge concerning uniqueness once the second part of Perry's account is in place.

(III) Self-notions: the output component

In the first part of this section (i), I describe Perry's candidate for being the role that self-notions play in action explanation, and I argue that self-notions do not play that role. In the second part (ii), I identify a central role that self-notions do seem to play in action explanation, but that Perry's theory is not equipped to acknowledge.

(i) *Where the action is not: the Special Actions hypothesis.*

In relation to action-explanation, Perry's claim is that there is a certain relationship – the *being the motivator of* relationship – that holds between self-notions and a certain set of types of action – types of action "whose success normally depends on facts about the agent".

Perry never says anything about how to understand the being the motivator of relationship. Intuitively though, it is whole attitudes (a desire, say), or perhaps even combinations of whole attitudes (a belief-desire pair, say) that motivate us to do things. So the most natural way of understanding what it is for a *notion* to motivate an action is in terms of the notion's being a *constituent* of a whole attitude that does some motivating. E.g. If my desire to hit Jack causes me to hit Jack, then my Jack-notion might be said to play a role in motivating me, in so far as it is a component of an attitude that motivates me.

The next question is *which types of actions* Perry has in mind. Which types of actions are such that their success "normally depends on facts about the agent"?[3]

If I attempt to ϕ, whether my attempt is successful depends precisely on whether I ϕ. That I ϕ is a fact about me, the agent of the action. At least in this sense, whatever exactly ϕ-ing is, its success will depend on facts about the agent. On the other hand, it would be uneconomical, and thus conversationally misleading, to say that self-notions are the motivators of actions whose success normally depends on facts about the agent, if what one actually thought was that self-notions are the motivators of *all* actions. So, it is natural to assume that the grammatically restricting phrase 'whose success normally depends on facts about the agent' is intended to effect a genuine restriction, i.e. to pick out a condition that some but not all actions meet.

Perry makes an effort to spell the condition out by giving examples. The first is taken from a scene in the Hitchcock film, *Spellbound*. In the scene, Leo G. Carroll, from whose point of view the action is seen, is initially pointing a gun at someone else, but then turns it in the direction of his own head.

[3] For brevity, I shall use 'action' as a name for action types. When talking about tokens of actions, I will use 'action-token'.

> Slowly, we see the hand holding the gun turn until the barrel of the gun is all that is visible on the screen. Then it fires. We know what Carroll has done, and to whom. He has killed someone, and the someone is him. The way Carroll held and fired the gun was a normally self-effecting way of killing someone. Of course, if Carroll had a head shaped like a donut, he could have shot the someone behind him. But normal people normally kill themselves when they shoot like that.
>
> This is only a particularly dramatic case of a whole class of actions. Imagine George and Barbara Bush seated across from each other at a boring dinner. Both know that the President is thirsty. Both may desire that he get a drink. The appropriate action for the President to take is the familiar one of reaching out and bringing the glass to his lips. That is an action that will succeed if the agent is thirsty. It is a normally self-dependent/directed/effecting action. (1990: 26-27)

Very roughly, the idea is that the relevant actions have the subject as their object. They *affect* the subject in certain ways (so Perry's 'effecting' should really be 'affecting' – the idea is that they *affect* the subject, not that they *effect* the subject, i.e. cause the subject to come into being). The actions' success depends on how things are with the subject in the specific sense that whether the actions are successful depends on whether the subject ends up being the *object* of a certain action -- e.g. being shot, having their thirst quenched. For economy of expression, but also to make the basic shape of Perry's idea as clear as possible, I shall refer to such actions as actions that are *on oneself*.

The phenomenon of acting on oneself, and not knowing it, yields a simple worry about Perry's theory. Here it is:

> One can think about oneself, without thinking a first-person thought about oneself. This is the most basic datum in this area of philosophy. But, the mere fact that this is one's situation need not prevent the thought one is having about oneself from leading to action. We can imagine a scenario in which I think of myself under the heading 'that dangerous rival, NN'. I might order a hit put out on NN. The order might be acted on and I, NN, might then be killed. In this case, I will have acted *on* myself. But this isn't the kind of action on myself that is likely to be especially closely tied to self-notions. Intuitively, the fact that the action is *on myself* in this case is irrelevant to its causal history. I am acting on myself *as though* I were acting on another. So, the causal history of the action – in particular, which notions play a motivating role – will be most similar to that of an action that really is on another, not to that of an action that is on myself, and that I *know* to be on myself.

I think this simple worry actually fully captures what is wrong with Perry's idea. Establishing that the worry cannot be replied to, however, will take a bit of argument.

First, notice that the worry simplistically assumed that 'acting on myself' is an extensional notion: the mere fact that I have acted on NN, and am identical to NN,

is assumed to entail that I have acted *on myself*. Someone might reject this assumption.

Rejecting it, however, is not a promising line for someone who likes Perry's idea. If acting *on oneself* is an intensional notion, it is, more particularly, a *de se* intensional notion. One *acts on oneself* – in the full-blooded intensional sense -- iff there is an object x such that (i) one acts on x (ii) one is identical to x and (iii) one's action on x is motivated by a de se attitude. If that is what acting on oneself comes to, then an account of self-notions that appealed to acting on oneself would be circular. Since I assume that we are interested in having, and that Perry is interested in giving, a non-circular account, I will continue to treat acting on oneself as an extensional notion. Perhaps the most plausible semantic hypothesis about the 'acting on' construction is that, sometimes, it is opaque, so that it expresses an intensional notion, and sometimes transparent, so that it expresses an extensional notion. If so, the uses of 'acting on' in this paper should always be read transparently.

The 'putting a hit out on NN' case that is the focus of the simple worry shows that self-notions aren't plausibly thought of as relevant to just any old case of acting *on oneself*. Perry's theory needs to restrict its focus to a narrower range of cases. The simplest way of achieving this restriction would be to appeal, implicitly or explicitly, to some de se notion. But, as already mentioned, that would make the account circular. Perry's account appeals to the notion of *normality* – his focus is not just any old actions that can happen to be on oneself, but actions that are *normally* on oneself. There is nothing particularly de se about the notion of normality. So, Perry's account eschews blatantly de se notions. Does it nevertheless achieve an appropriate restriction?

One can see why it might initially seem to. The second action Perry mentions in the passage quoted above is something like:

> Raising a glass of water to one's lips.

This action contrasts with, for example:

> Putting a hit out on NN.

A token of the second action might *happen* to be on oneself. It will be just in case it is a token action of NN's. But it will not be *normally* on oneself.

Consider, however, the following action:

> Putting a hit out *on oneself*.

This is an action that is necessarily – a fortiori, normally – on oneself. But a very significant proportion of the cases in which the action is performed will be cases in which the subject fails to know it is themselves they are acting on (e.g. they want to put a hit out on someone they conceived of as 'that dangerous rival', and that person happens to be themselves). In such cases, there is no reason to expect self-notions to be involved in motivating the action.

One might consider offering the following, more complicated theory on Perry's behalf:

> Self-notions are the motivators of *basic* actions that are normally on oneself.

A *basic action* is an action that one can perform *just like that*, i.e. without performing some other action that stands to that action in the relation of means.[4] 'Raising a glass of water to one's lips' very plausibly refers to a basic action. 'Putting a hit out on oneself' surely doesn't. So, one might think, something like the idea of a basic action is already playing some role in controlling Perry's choice of example, and the amended theory simply involves making explicit something Perry implicitly assumed. Appealing to basicness is OK since, just like the notion of normality, basicness is not a de se notion.[5]

Adding basicness to normality ultimately fails to help, however. Notice first that one can perform even *basic* actions on oneself without thinking of oneself in a de se way. For example, I might wake up after an accident and momentarily see a hand that I take to be someone else's but that is in fact my own. I might form the intention to touch *that hand*, and actually touch it. This would be a token of a *basic* action on myself that is not motivated by my self-notion.[6]

Admittedly, it would not be a token of a basic action that is *normally* on myself. To come up with that, we need a slightly more complicated set up. Suppose I live on a planet that is full of mirrors, in which each inhabitant regularly sees themselves. Despite this, I consistently fail to realize that it is me I am seeing. Whenever I see myself, I fire paint balls at myself, conceived of under the guise 'that threatening person', and I flee. The physics of the world is such that the balls end up reaching me, always so diminished in momentum as to escape my notice on landing. 'Firing at oneself' is a basic action of mine that is normally on myself. But, it is not motivated by my self-notion. It is motivated by a perceptual demonstrative notion, expressed by 'that threatening person'.

One might suggest that, if the situation really is as described – if my perceptual demonstrative notion is normally leading me to perform actions that have myself as their object – then it will somehow take on the character of a self-notion. But, on the face of it, this is the wrong verdict. I may still do very well in forming other first-person beliefs about myself --- e.g. "I fire a lot of paint-balls", "I am unaccountably covered in tiny specks of paint'. At some point, I may realize

---

[4] See Hornsby 1980.

[5] Perry suggested this restriction in a conference at which I presented this objection.

[6] Wittgenstein (1958) famously discusses this possibility of seeing a part of one one's body without recognizing it as one's own. Sacks (1985) discusses subjects who see parts of their body without recognizing them as their own, and who moreover act on those body parts as though they were not their own – e.g. patients who throw their own arms out of bed, and find themselves lying on the floor. Clearly, the waking-up-from-an-accident cases, involving non-delusional subjects, can equally involve action.

what's going on. The natural way of describing what I come to know is that *that person* is me. Knowing that is different from knowing the trivial fact that *I* am me.

Since the example involves a subject in a world with weird physics it might be worthwhile comparing it with the example Perry has of a subject with weird physiology -- the subject with a head shaped like a donut. Perry's point here is that that the subject is not a problem for his theory because the ways of acting his theory appeals to is only meant to be *normally* self-effecting, not *necessarily* self-effecting. By contrast, the subject in a world with weird physics is a counterexample. 'Shooting at oneself' is necessarily, a fortiori normally, an action on oneself. But, as the weird physics case brings out, it need not be motivated by a self-notion.

Perry's theory is naturally read as trying to use the extensional concept of normal self-directedness to simulate the intensional concept of an action that is performed on oneself *as oneself*. The need to avoid circularity makes this attempt understandable. It turns out not to work though. The cases in which one acts on oneself *as oneself* are not the same as the cases in which one performs actions that are normally on oneself.

(ii) *Where the action is: The Just Acting Hypothesis.*

Suppose one has accepted that the notion of self-directedness cannot be used to specify a class of actions that is particularly closely related to self-notions. One option is to try to find some different way of demarcating a subset of actions that are especially closely related to self-notions. The more radical option is to abandon the assumption that the role of self-notions in action explanation consists, more particularly, in a relationship to a particular subset of actions. That is what I propose.[7] Here is my hypothesis:

> Whenever one ϕ-s, one's action will be motivated by an attitude that involves a self-notion (*The Just Acting Hypothesis*).

A way of defending this hypothesis is to point out that, whatever particular action ϕ-ing is, one's intending or desiring that *x* ϕ isn't enough to get one to ϕ, even if one is x. I might, for example, desire, or even intend, that NN flee a bear. But if I assume that NN is someone other than me, it's hard to see how having these attitudes is going to get me to flee a bear. I need to intend that *I* flee, or at least to intend *to flee*. And that seems to involve self-notions.

It's worth pointing out immediately that, if *The Just Acting Hypothesis* is correct, the way I summarized the problem for Perry's theory in the simply worry described above was, strictly speaking, false. Suppose I put hit out on myself, conceiving of myself as NN. I said it was a problem for Perry's theory that, in

[7]Gjelsvik (this volume) proposes something similar. On Gjelsvik's view, as I read it, states of practical knowledge with de se contents – e.g. "I am turning left" – are *identified* with the actions – e.g. turning left– they involve practical knowledge of. So, de se states play a special role in constituting actions – every action, not just some actions. By contrast, my view is that de se states play a special role in causing actions – every action, not just some.

such a case, my self-notion is not relevant to my action. In fact, if *The Just Acting Hypothesis* is true, it will be relevant. It is just that my self-notion is not relevant in the way that it would be in a case in which I put a hit out on myself, conceiving of myself as myself, and therefore intend to put a hit out *on me*. That is, it is not relevant in virtue of specifying the object of my action. And it is this role for the self-notion in action-explanation that Perry's theory is fixated on. If *The Just Acting Hypothesis* is correct, this is not the only role for the self-notion in action explanation.

It might be objected to the initial argument for *The Just Acting Hypothesis* in the last but one paragraph that the difference between 'intending that I flee' and 'intending to flee' is significant. In particular, the idea would be that 'intending to flee' picks out a different kind of intention from 'intending that I flee', one that does not involve a self-notion because it is only *implicitly* about the subject.[8] Since intending to flee is surely sufficient for action, this undermines *The Just Acting Hypothesis.*

One response would be to concede the point about intentions, and retreat to mental states whose status as involving self-notions is less controversial. For example, one might argue that even if intending to flee a bear does not involve a self-notion, one will not form that intention unless one has some relevant belief that involves a self-notion – e.g. 'I am being pursued by a bear'.

Doing this has costs. It threatens to lose generality. It isn't obvious that whenever one forms an intention to ϕ, that must be because one has first formed some *belief*. What if I form an intention to click my fingers just for its own sake? It isn't obvious that *any belief* must play a role in explaining why I form the intention that I do.

The maneuver is also unsuccessful in its own terms. The view being considered about intentions-to – that they are not explicitly about the subject – is also a viable option about first-person beliefs. [9]

The more convincing reply is that there is an intuitive and robust distinction between first person attitudes and non first-person attitudes, and that an intention to ϕ falls on the first-person side of this distinction. There is some kind of tension or irrationality involved in intending to ϕ, while believing that *I* cannot ϕ, or believing that *I* ought not ϕ. There isn't any kind of tension or irrationality involved in intending to ϕ, while believing that *NN* cannot ϕ or believing that *NN* ought not to ϕ, unless one realizes that one is NN. That warrants putting intentions-to in the same category as paradigm first-person attitudes, not the same category as paradigm non first-person attitudes. There may be important differences among the different members of this broad category– e.g. in whether they are explicitly or only implicitly about the subject. But it is the conceptual role of this broad category we are interested in. If we were to stipulate that an

[8] See Rumfitt, I (1994)

[9] The view is famously defended in Lewis (1979).

attitude only involves a self-notion if it is *explicitly* about the subject, then it would turn out that the conceptual role of self-notions is not the quite same thing as we are interested in.

In searching for a proper subset of actions to which self-notions are especially relevant, Perry's misses the possibility that self-notions are unusual precisely in being involved *whenever* one acts. Perry's most famous example of first-person action explanation – a case in which someone performs the action of running away on seeing a bear – does not, in fact, involve someone performing an action *on themselves*, and for that reason it is not in the remit of Perry's official account of the role of self-notions.[10] Nevertheless, Perry seemed right first time to present it as a perfectly good example of first-person action explanation. Perry's official theory focuses on the fact that the subjects can act on themselves. This approach turns out to founder on the datum that one can act on oneself without knowing it. The *Acting Tout Court Hypothesis* I propose instead focuses on the fact that the subject of thought is the *author* of action. The special work self-notions do isn't in getting me to act *on me*. It is in getting *me* to act. There is an echo here of the Wittgensteinian distinction between uses of "I" 'as object' and 'as subject'.[11] Wittgenstein had the idea that uses of "I" 'as object' were less fundamental that uses of "I" 'as subject'. His distinction was oriented towards the input component of first-person thought. It is sometimes suggested that there ought to be an analogous distinction that is oriented towards the output component. On my view, this is right, though it's more natural to label the distinction as one between uses of "I" 'as object' and 'as author'. Self-notions *can* be used to specify the object of an action.[12] But they *must* specify its author.

(IV) Top-Down Distortion

One can imagine accounts of the role of a particular notion being influenced by two fairly different kinds of consideration. One kind of consideration is *bottom up.* For example, being in pain seems to justify the thought 'I am in pain' in a way that it does not justify the thought 'NN is in pain'. This is a bit of data that might end up feeding into an account of self-notions. The other kind of consideration is *top-down* – i.e. general assumptions about what accounts of any notion's conceptual role must be like. I want to argue that Perry's account is a victim of top down distortion. In the last two sections, I identified some ways in which his

[10] The bear example appears in Perry 1979. Interestingly Perry does not comment in his post-1990 papers on the role of self-notions on how these relate to his earlier, ground-breaking papers in 1977 and 1979. Perhaps he regards the later papers as developing insights that were merely gestured at in the earlier papers. My suggestion here is that there are insights in the earlier papers that are contradicted by what he says in the later papers.

[11] See Wittgenstein (1958: 66-7)

[12] At least they can for most subjects. It's not clear that they can for all. Suppose one's brain is envatted and one is remotely controlling, in rapid succession, a variety of bodies located thousands of miles away. Arguably, none of these bodies is durably enough connected with oneself to count as 'one's body' in such a way as to justify the claim that in acting on one of them one would be acting on oneself. Or suppose that one's brain is inside one's body in the normal way, but one's body is restricted enough in its range of movements that it can be used to act on things in its environment, but not on itself. By contrast, there do not seem to be metaphysically possible subjects who act without self-notions specifiying the author of their actions.

theory seems to go wrong. In this section, I want to trace his theory going in wrong in these ways to three general assumptions about conceptual roles -- *Symmetry, Generalizability,* and *Uniqueness*. Part of the interest here is diagnostic in relation to Perry's account. But the larger aim is to point in the direction of a better account of first-person thought's conceptual role, one that avoids top-down distortion.

> *(i) Symmetry*: The input and the output components of a concept's role must be symmetric.

Symmetry is vague – symmetry comes in degrees. But that doesn't mean it is lacking in content. Suppose the canonical conceptual role for AND is the following.

> Input component: If one judges that p and one judges that q, one will be disposed to judge p and q.
>
> Output component: if one judges that p and q, one will be disposed to judge that p, and one will be disposed to judge that q.

Each of these transitions in thought is the mirror image of the other. So, if these transitions constitute the conceptual role for AND, AND clearly satisfies *Symmetry.*

There is good reason to think that *Symmetry* must hold for a concept of a logical constant like AND. If it did not, AND would be non-conservative. That is, introducing the concept AND would enable to establish contents *not* involving AND which one previously had been unable to establish.

However, this reason for insisting that AND must satisfy *Symmetry* fails to generalize to self-notions. In their case, neither the input nor the output component plausibly consists in an inferential transition. So, there is no risk of a *bad* inferential transition being permitted by the role of the concept. There might be a more complicated argument for *Symmetry* that generalizes beyond concepts whole roles consist in inferential transitions. But it isn't obvious what that is. So, we have no positive reason to endorse *Symmetry*. I now turn to two points at which endorsing it could do damage in an account of self-notions.

(i) An account of the role of self-notions that includes *The Just Acting Hypothesis* is unlikely to satisfy *Symmetry*. *The Just Acting Hpyothesis* says that whenever one's acts, one's action will be motivated by an attitude that involves a self-notion. There is no analogue of this claim on the input side that has any plausibility. One can learn things about the world without learning anything about oneself. So, accepting *Symmetry* will tend to make one miss *The Just Acting Hypothesis.*

(ii) I agreed with Perry that there is an interesting connection between self-notions and ways of gaining information that normally have the self as their object. I disagreed with him about there being any interesting connection between self-notions and types of actions that normally have the self as their object. If one accepts *Symmetry* one is likely to think that one must agree with him on both counts or neither. It is plausible that Perry's having the insight that a

good theory of the input component ought to focus on normally self-informative ways of gaining information that are is part of what led him to concoct a bad theory of the output component that focused on actions that normally have the self as their object. *Symmetry* is a linking thought between the good and the bad here.

> (ii) *Generalizability*: An account of the role of first-person thought must be constructed from materials that enable it to generalize far beyond the case of first-person thought.

Unlike *Symmetry*, *Generalizability* is a claim Perry more or less *explicitly* aims to satisfy. His account of the role of self-notions is accompanied by more programmatic suggestions about how the account can be extended to a very wide range of notions. The basic idea is that the world is full of what he calls "epistemic/pragmatic" relations:

> I shall call relations between an agent and another object – including places, material objects and other persons – that support… special ways of knowing and acting, "epistemic/pragmatic" relations. The relation of being in, a relation between people and places, is an epistemic/pragmatic relation. There are many others. There are special ways to know about the material objects and people in front of one (open your eyes and look, reach out and touch), and special ways of dealing with them. There are special ways of knowing about the person who wrote the article you are reading, and special ways of communicating with them. There are special ways to know what a person is saying when they have called you on the phone (listen to the sounds coming out of the ear piece) and special ways of saying things to them (speak into the mouthpiece). Where R is an epistemic/pragmatic relation, we may speak of "normally R-informative ways of perceiving" and "normally R-directed/dependent/effecting ways of acting". (1990: 24)

The following schema capture how these epistemic-pragmatic relations figure in his account of the role of notions *in general*:

> The informational role of an R-notion is to serve as the normal repository for information gained in normally R-informative ways, and as the normal motivator for normally R-effecting and R-dependent actions (1990: 24).

In the case of self-notions, under discussion in this paper and developed by Perry in far more detail than any of the other potential instances of the schema, relation R is the relation of being identical to.[13]

It is hard to see how *The Just Acting Hypothesis* could fit it in with *Generalizability*. The hypothesis says that, whenever you act, you need a first-person thought. The only way of generalizing it would be to say that it is also true of many other kinds

[13] Self-notions are probably the most plausible-seeming case for the account, which is why Perry focuses on them. Different instances of the schema will have different extra problems. Consider a past-tense 'then'-notion. According to the schema, it should be the motivator of actions that affect the past. But this would involve backwards causation.

of thought that whenever you act, you need a thought of that kind. That has little plausibility (beyond a very few other cases – e.g. 'now'-thoughts).

By contrast, if one focused on the way the self-notion can specifies the object of an action that might seem more likely to generalize. A place or a material object other than myself cannot be the author of my actions. But they could be the object of my action. So, if one accepts *Generalizability*, one has a reason to focus on the role self-notions play in specifying the objects of certain actions – a line of thought that Perry clearly follows but that I argue proves unprofitable in the end.

Perry does not offer any motivation for *Generalizability*. Clearly, more general accounts of anything are better than less general accounts, *ceteris paribus*. But why couldn't the most illuminating account of the role of self-notions be one that failed to generalize very far beyond self-notions?

Here is a possible answer. The ultimate point of describing a concept's role is to contribute to the goal of naturalizing intentionality. If that is one's goal then having an account of *one* concept's role that failed to generalize to *other* concepts would be pretty useless. All intentionality is equally need of naturalization. Perry's framing of his account of first-person person thought as an account of self-notions – cognitive particulars *in one's brain* – makes him sound well disposed to naturalism.

It isn't clear that Perry could appeal to this motivation, since it isn't clear that the resources that figure in his theory are naturalistically acceptable. For example, the theory includes an unreduced appeal to special 'ways of knowing' or 'ways of gaining information'. These seem intentional.

Leaving Perry aside, the naturalistic motivation is underwhelming, since there are good reasons to be interested in the role of first-person thought that have nothing to do with project of naturalizing intentionality. At the beginning of this paper, I began by raising the question of why first-person thoughts are more than just a convenience. A plausible first stab at an answer to that question is to mention that they have a special conceptual role. But a more substantive answer will say what the role is. This kind of motivation provides no support for *Generalizability*.

There is also a positive consideration against *Generalizability*. The examples of action-explanation that Castaneda and Perry made famous are usually taken to support the idea that some indexical concepts – e.g. self-notions or now-notions – are *especially* relevant to action explanation.[14] If it turns out our best description of the role self-notions play in action explanation is just one instance of a much more general schema, then it turns out that they are not *especially* relevant to action explanation. Conversely, if one thinks that that the original Casteneda-Perry thought about their being *especially* important is an insight, one should expect *Generalizability* to fail.

[14] See Castaneda 1968, Perry 1977 and 1979.

> (iii) *Uniqueness:* An account of the role of first-person thought must capture something unique to it—i.e. a role that it has and that no other kind of thought has.

Perry's theory, on the face of it, is well designed to satisfy *Uniqueness.* Every notion will be associated with a relationship *R* that describes the relationship between the thinker and referent of the notion, as used on a given occasion. Corresponding to these relationships, there will be different R-effecting actions and R-informative ways of gaining information. So, the different roles associated with different notion look as though they will be different.

In fact, things turn out to be more complicated. I pointed out above that 'normally self-informative ways of gaining information' and 'normally current time informative ways of gaining information', although they clearly differ in sense, might turn out to refer to *the same* ways of gaining information. Perception, introspection and internal bodily awareness all seem to be *both* normally self-informative and normally current-time informative. So, although Perry's account looks to be aiming to be satisfying *Uniqueness,* it isn't clear that it pulls it off.

Is there a good argument for *Uniqueness*? Again, one might try to appeal to the project of naturalizing intentionality. If one is trying to say something like: this particular cognitive particular in the brain, identified in naturalistically acceptable terms, is the vehicle for a self-notion because it plays role *r ,* it had better be the case that *r* is associated *only* with self-notions, rather than being the common property of, for example, self-notions and now-notions. But, if that isn't one's motivation, it isn't clear why one should endorse *Uniqueness.*

Conclusion:

The focus of this paper has been the question 'What is the conceptual role of first-person thought?' I've first of all suggested that it would be a mistake to treat this question as already having been answered, in the way that one does if one focuses on addressing a question that can only reasonably be addressed once one has a rough idea of what the conceptual role of first-person thought is -- e.g. the question 'What is the relationship between the conceptual role of first-person thought and its pattern of reference? I've also made some positive suggestions about the conceptual role of first-person thought, in particular defending the *The Just Acting Hypothesis*. More generally, I've argued that, in giving an account of the conceptual role of first-person thought, there is significant risk of top-down distortion. Perry's account is offered as a clear example of an account that goes wrong because it accepts high-level assumptions about what an account of a conceptual role must look like that are only motivated for certain concepts (e.g. *Symmetry*, which is only motivated for concepts of logical constants) or relative to certain motivations for giving an account of a conceptual role (e.g. *Generalizability* and *Uniqueness*, which are compelling only if one is invested in the project of trying to naturalize intentionality).

References:

Campbell, J (2004) 'What is it to know what "I" refers to', *The Monist* 87 (2004), 206-218.

Cappelen, H and Dever, J (2013) *The Inessential Indexical: on the Philosophical Insignificance of Perspective and the First Person.* Oxford: Oxford University Press.

Castaneda, H.N (1968). 'On the Logic of Attributions of Self-Knowledge to Others', in *Journal of Philosophy*, 65, 439-56.

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

Gjelsvik, Olav (2016) 'Indexicals: what they are essential for' *Inquiry*

Hornsby, J (1980). *Actions.* Routledge and Kegan Paul.

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

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

Perry, J (1977). 'Frege on Demonstratives', in *The Philosophical Review* 86 (4): 474-497.

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

Perry, J (1990). 'Self-Notions', in *Logos*: 17-31.

Perry, J (2010). 'Selves and self-concepts'. In Joseph Keim Campbell, Michael O'

Rourke, and Harry S. Silverstein (eds.) *Time and Identity* (Topics in Contemporary Philosophy), MIT Press: 229-248.

Perry, J (2011). In 'On Knowing Your Self.' In Shaun Gallagher (ed.), *The Oxford Handbook of the Self*. Oxford: Oxford University Press, 2011.

Perry, J (2012). 'Thinking About the Self' .In JeeLoo Liu and John Perry, (eds.), *Self and Conscioiusness*. Cambridge: Cambridge University Press.

Recanati, F (2007) *Perspectival Thought: A Plea for (Moderate) Relativism.* Oxford: Oxford University Press.

Rumfitt, I (1994) 'Frege's theory of predication: An elaboration and defense, with some new applications' *Philosophical Review* 103 (4): 599-637

Sacks, O. (1985) *The Man who mistook his wife for a hat. Gerald Duckworth*

Wittgenstein, L. (1958) *The Blue and Brown Books*. New York: Harper and Row.