Consciousness and the Self: New Essays, ed. J. Liu and J. Perry (Cambridge University Press)

Knowing What I Want[1]

Alex Byrne

> As it is silly to ask somebody,'How do you know you are in pain?' it is equally foolish to ask, 'How do you know that you want to go to the movies?'
>
> Vendler, *Res Cogitans*

Knowing that one wants to go to the movies is an example of self-knowledge, knowledge of one's mental states. It may be foolish to ask the man on the Clapham Omnibus how he knows what he wants, but the question is nonetheless important — albeit neglected by epistemologists. This paper attempts an answer.

Before getting to that, the familiar claim that we enjoy "privileged access" to our mental states needs untwining (section 1). A sketch of a theory of knowledge of one's beliefs that has received some attention in the recent literature (section 2), and the case for extending that account to self-knowledge in general (section 3), sets the stage for our answer to the main question (section 4).

1. A "Twofold Privileged Access"

The term "privileged access" is due to Gilbert Ryle; on the Cartesian view that he is concerned to attack, "[a] mind has a twofold Privileged Access to its own doings" (Ryle 1949: 148). The first kind of privileged access is that:

[1] Thanks to audiences at the University of Oxford and Cal State Fullerton, and to Lauren Ashwell, Caspar Hare, Richard Holton, Rae Langton, JeeLoo Liu, Julia Markovits, Eric Schwitzgebel, Ralph Wedgwood, Tim Williamson, and Steve Yablo.

> (1) ...a mind cannot help being constantly aware of all the supposed occupants of its private stage, and (2) ...it can also deliberately scrutinize by a species of non-sensuous perception at least some of its own states and operations (148).

And the second kind is that:

> both this constant awareness (generally called 'consciousness'), and this non-sensuous inner perception (generally called 'introspection') [are] exempt from error. (149)

(1) in the first quotation together with the first conjunct of the second quotation basically amount to the claim that one's mental states are *self-intimating*: if one is in a mental state M, one knows (or believes) that one is in M.

Self-intimation is extremely implausible, at least with respect to the mental states that will be under discussion in what follows. Here is a perfectly ordinary *prima facie* example of believing that *p* (and possessing the relevant concepts) without knowing, or even being in a position to know, that one believes that *p*. Pam is now not in a position to retrieve the name of her new officemate (it is on 'the tip of her tongue', say) and so is not in a position to verbally self-ascribe the belief that her officemate's name is 'Andy.' Nothing else in her behavior, we may suppose, indicates that she believes that she has this belief. But she nonetheless does believe now that her officemate's name is 'Andy,' because otherwise there would be no explanation of why she recalls the name later when taking the train home.[2]

---

[2] And if Pam *does* believe that she believes that her officemate's name is 'Andy,' the question whether she believes *that* arises, and so on through progressively more iterations. This regress stops somewhere, presumably.

Of more interest is (2) in the first quotation, the claim that one employs a kind of "non-sensuous perception" to find out about one's own mental life. Non-sensuous perception is supposed to work only in one's own case: according to Ryle's opponent, "I cannot introspectively observe… the workings of your mind" (Ryle 1949: 149). A more general version of (2), then, is the claim that one knows about one's mental life in a way that one cannot know about another's mental life. That is, one has a special method or way of knowing that one believes that *The Searchers* is playing at the Orpheum, that one wants to go to the movies, and so on, which one cannot use to discover the mental states of someone *else*. Since such a first-person method need not be epistemically privileged or authoritative, 'privileged access' is not ideal terminology. Instead, let us put the generalized (2) by saying that we have *peculiar* access to our mental states.

Ryle's second kind of "Privileged Access," "exemption from error," is evidently better-named. Infallibility sets the bar too high, though: a weaker and more useful claim is that our beliefs about our own mental states, arrived at by typical means, are more likely to amount to knowledge than the corresponding beliefs about others' mental states and the corresponding beliefs about one's environment. (The latter comparison has clear application only for some mental states, paradigmatically belief.) Retaining Ryle's terminology, let us put this weaker claim by saying that we have *privileged* access to our mental states.

A number of authors (for instance, Alston 1971; McKinsey 1991; Moran 2001: 12-3), presumably following Ryle, use 'privileged access' for what is described in our preferred terminology as privileged *and* peculiar access. Whatever the labels, it is important to keep the two sorts of access separate.

The distinction between privileged and peculiar access is one thing; the claim that we actually have one or both sorts of access is another. Let us briefly review some evidence.

1.1. Privileged access

Consider Jim, sitting in his office cubicle. Jim believes that his pen looks black to him; that he wants a cup of tea; that he feels a dull pain in his knee; that he intends to answer all his emails today; that he is thinking about paperclips; that he believes that it is raining. Jim also has various equally humdrum beliefs about his environment: that it is raining, that his pen is black, and so on. Furthermore, he has some opinions about the psychology of his officemate Pam. He believes that her pen looks green to her; that she wants a cup of coffee; that her elbow feels itchy; that she is thinking about him; that she believes that it is raining.

In an ordinary situation of this kind, it is natural to think that Jim's beliefs about his current mental states are, by and large, more epistemically secure than his corresponding beliefs about this officemate Pam and his corresponding beliefs about his environment.

Take Jim's belief that he believes that it is raining, for example. It is easy to add details to the story so that Jim fails to know *that it is raining*; it is not so clear how to add details so that Jim fails to know that he *believes* that it is raining. Perhaps Jim believes that it is raining because Pam came in carrying a wet umbrella, but the rain stopped an hour ago. Jim is wrong about the rain, but he still knows that he *believes* that it is raining — this knowledge will be manifest from what he says and does.

Now contrast Jim's belief that he believes that it is raining with his belief that Pam believes that it is raining. Again, it is easy to add details to the story so that Jim fails to know that Pam believes that it is raining. Perhaps Jim believes that Pam believes that it is raining because he entered the office wearing a visibly wet raincoat. Yet Pam might well not have noticed that the raincoat was wet, or she might have noticed it but failed to draw the obvious conclusion.

Similar remarks go for Jim's belief that he wants a cup of tea, which can be contrasted with Jim's belief that Pam wants a cup of coffee. Now it may well be that, in general, beliefs about one's own desires are somewhat less secure than beliefs about one's own beliefs, or beliefs about how things look. This is more plausible with other examples, say Jim's belief that he wants to be the CEO of Dunder Mifflin Paper Company, Inc., or wants to forever remain single — it would not be particularly unusual to question whether Jim really has these particular ambitions. And perhaps, in the ordinary circumstances of the office, Jim might even be wrong about his desire for tea. Still, Jim's claim that he wants tea would usually be treated as pretty much unimpeachable, whereas his claim that Pam wants coffee is obviously fallible. (Jim's evidence points in that direction: Pam normally has coffee at this time, and is heading to the office kitchen. However, she drank her coffee earlier, and now wants a chocolate biscuit.) And treating Jim as authoritative about his own desires has nothing, or not much, to do with politeness or convention. Jim earns his authority by his subsequent behavior: Jim will drink an available cup of tea and be visibly satisfied.

The precise extent and strength of privileged access is disputable; the fact of it can hardly be denied.[3]

1.2. Peculiar access

Peculiar access is equally apparent. The importance of "third-person" evidence about one's mental life can easily be overlooked, but it is clear that one does not rely on such sources alone. Quietly sitting in his cubicle, Jim can know that he believes that its raining and that he wants a cup of tea. No third-person or behavioral evidence is needed. To know that Pam wants a coffee requires a different sort of investigation — asking her, observing what she does, and so forth.

It is often claimed that one knows one's mind "directly," or "without evidence." (For the former see, e.g., Ayer 1959: 58; for the latter, see, e.g., Davidson 1991: 205.) If that is right, and if one know others' minds "indirectly," or "with evidence," then this is what peculiar access consists in — at least in part. But such claims should be made at the end of enquiry, not at the beginning.

Sometimes peculiar access is glossed by saying that self-knowledge is "a priori" (e.g., McKinsey 1991, Boghossian 1997). This should be resisted — certainly at the beginning of enquiry, and probably at the end. One leading theory of self-knowledge classifies it as a variety of *perceptual* knowledge, in many respects like our perceptual

---

[3] Schwitzgebel (2011) argues that there is much about our own mental lives that we don't know, or that is difficult for us to find out, for instance the vividness of one's mental imagery, whether one has sexist attitudes, and so forth. His treatment of individual examples may be questioned, but his overall argument is an important corrective to the tendency to think of the mind as an internal stage entirely open to the subject's view. However, too much emphasis on this point can lead to the opposite vice, of thinking that self-knowledge poses no especially challenging set of epistemological problems.

knowledge of our environment. "The *Perception of the Operations of our own Minds* within us," according to Locke, "is very like [the perception of "External Material things"], and might properly enough be call'd internal Sense" (Locke 1689/1975: 105). On this *inner-sense theory* (Armstrong 1968: 95; see also Lycan 1987: ch. 6, Nichols and Stich 2003: 160-4), we have an internal "scanner" specialized for the detection of our mental states. No doubt the hypothesized inner sense is not much like our outer senses — recall that Ryle characterizes it as "non-sensuous perception" — but it is surely unhelpful to classify its deliverances with our knowledge of mathematics and logic.[4]

1.3. The independence of privileged and peculiar access

It is important to distinguish privileged and peculiar access because they can come apart in both directions. Hence one can find theorists who deny that we have one kind of access while affirming that we have the other (for examples, see Byrne 2005: 81). As the previous two sections suggest, this extreme claim is not credible, but a more restricted version is actually correct. Privileged and peculiar access do not perfectly coincide: in particular, there are many ordinary cases of the latter without the former.

For instance, the epistemic security of self-ascriptions of certain emotions or moods is at the very least nothing to write home about. One may have peculiar access to the fact that one is depressed or anxious, but here the behaviorist greeting — "You're fine! How am I?" — is not much of a joke, being closer to ordinary wisdom.

---

[4] See McFetridge 1990: 221-2, Davies 1998: 323. The classification of (much) self-knowledge as a priori has its roots in Kant's definition of a priori knowledge as "knowledge absolutely independent of all experience" (Kant 1787/1933: B3; see McFetridge 1990: 225 and McGinn 1975/6: 203).

Factive mental states, like *knowing that Ford directed* The Searchers and *remembering that the Orpheum closed down last week*, provide further examples.[5] Since knowing that Ford directed *The Searchers* entails that Ford directed *The Searchers*, but not conversely, it is easier to know the latter fact than to know that one knows it.[6] The belief that one knows that Ford directed *The Searchers* is *less* likely to amount to knowledge than the belief that Ford directed *The Searchers*. Yet one has peculiar access to the fact that one knows that Ford directed *The Searchers*, just as one has peculiar access to the fact that one believes this proposition. Jim knows that Pam knows that Ford directed *The Searchers* because (say) he knows she is a movie buff and such people generally know basic facts about John Ford. But in order to know that he knows that Ford directed *The Searchers*, Jim need not appeal to this kind of evidence about himself.

2. Belief and BEL

How does one know what one believes? Evans suggested an answer: "in making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward — upon the world" (1982: 225). Different ways of elaborating Evans' telegraphic remarks have been proposed by Gordon (1995), Gallois (1996), Moran (2001), Fernández (2005), and Byrne (2005). In brief, here is the proposal in Byrne 2005.

First, a small amount of technical apparatus. Say that an *epistemic rule* is a conditional of the following form:

R If conditions C obtain, believe that *p*.

---

[5] For a defense of the claim that knowing is a mental state, see Williamson 1995.

[6] With the assumption that the "KK principle" (if one knows that *p*, one knows that one knows that *p*) is false.

An example is:

WEATHER If the clouds are dark grey, believe that it will rain soon.

One *follows* WEATHER on a particular occasion iff one believes that it will rain soon because one recognizes that the clouds are dark grey, where the 'because' is intended to mark the kind of causal connection characteristic of inference or reasoning. In general, S follows rule R on a particular occasion iff on that occasion S believes that *p* because she recognizes that conditions C obtain, which implies (a) S knows that conditions C obtain, which in turn implies (b) conditions C obtain, and (c) S believes that *p*.

Following WEATHER in typical circumstances tends to produce knowledge about impending rain; rules that are knowledge-conducive are *good* rules.

Now we can put an Evans-inspired proposal for the epistemology of belief as follows. Knowledge of one's beliefs can be obtained by following the rule:

BEL If *p*, believe that you believe that *p*.[7]

But is BEL a *good* rule, as the proposal implies? It might seem not, because its third-person counterpart is plainly a *bad* rule:

BEL-3 If *p*, believe that Pam believes that *p*.

It would be quite a stretch, for instance, to reason from the fact that I have a quarter in my pocket, or that the Philadelphia Phillies were orginally known as the 'Quakers,' to the conclusion that Pam believes these things — most likely she doesn't. Following BEL-3 will thus tend to produce false and unjustified beliefs. Isn't BEL just as dodgy?

---

[7] 'You' refers to the rule-follower; tenses are to be interpreted so that the time the rule is followed counts as the present.

Fortunately not. BEL, unlike BEL-3, is *self-verifying*: if one follows it one's second-order belief is true.[8] In this respect it is an even better rule than WEATHER — that rule is good, but following it does not guarantee that one will have a true belief about the rain.

The virtues of BEL do not stop there. Say that S *tries* to follow R iff S believes that *p* because she *believes* that conditions C obtain; hence *trying to follow R* is weaker than *following R*.[9] Trying to follow WEATHER is not truth-conducive, hence not knowledge-conducive. In contrast, trying to follow BEL is maximally truth-conducive: if one tries to follow it, one's second-order belief is true. We can put this by saying that BEL is *strongly* self-verifying. (Self-verification is thus weaker than strong self-verification.)

The Evans-inspired proposal of a few paragraphs back only covers cases where one's belief that *p* amounts to knowledge, but of course one may know that one believes that *p* even though one's belief that *p* is false, or anyway does not amount to knowledge. In the terminology just introduced, the fully general proposal for the epistemology of belief is this:

BELIEF Knowledge of one's beliefs can be, and typically is, obtained by trying to follow BEL.

BELIEF implies that *trying* to follow BEL is knowledge-conducive — that BEL is a *very* good rule.

---

[8] A qualification: since inference takes time, it is not impossible that one's belief in the premise vanishes when one reaches the conclusion. Since the inference is as simple and as short at they come, this qualification is of little significance.

[9] Assuming that '…because S knows P' entails '…because S believes P.'

Now true beliefs, or even true beliefs that could not easily have been false, are not thereby knowledge. So the fact that BEL is strongly self-verifying does not *entail* that it is a very good rule. It does, however, rebut an obvious objection. And given the intuitive plausibility of Evans' claim that we know what we believe by "directing our eyes outward," the burden of proof is on the critics. BELIEF is at least a defensible working hypothesis.

BELIEF, if true, offers a satisfying explanation of both privileged and peculiar access. Privileged access is explained because BEL is strongly self-verifying. Peculiar access is explained because the method involved only works in one's own case: third person rules like BEL-3 are bad.

For the sake of the argument, assume that BELIEF is true. Is there any need to press on, and try to force all self-knowledge — a fortiori, knowledge of one's desires — into this rough mold?

3. Economy and unification

In answering that question, two distinctions will be helpful. To introduce the first, consider our knowledge of metaphysical modality. On one (popular) view, it requires a special epistemic capacity of modal intuition. On the alternative Williamsonian picture, it requires nothing more than our "general cognitive ability to handle counterfactual conditionals" (Williamson 2004: 13), such as 'If it had rained I would have taken my umbrella.' The Williamsonian view is an *economical* account of our knowledge of metaphysical modality: all it takes are epistemic capacities required for other domains. The popular view, on the other hand, is *extravagant*: knowledge of metaphysical modality needs something extra.

A similar "economical-extravagant" distinction can be drawn for self-knowledge. Let us say that a theory of self-knowledge is *economical* just in case it purports to explain self-knowledge solely in terms of epistemic capacities and abilities that are needed for knowledge of other subject matters; otherwise it is *extravagant*. A behaviorist account of self-knowledge is economical: the capacities for *self*-knowledge are precisely the capacities for knowledge of the minds of *others*. The theory defended in Shoemaker 1994 is also economical: here the relevant capacities are "normal intelligence, rationality, and conceptual capacity" (Shoemaker 1994: 236). On the other hand, the inner-sense theory (see section 1.2 above) is *extravagant*: the organs of outer perception, our general rational capacity, and so forth, do not account for all our self-knowledge. For that, an additional mechanism, an "internal scanner," is needed.

The second distinction is between *unified* and *disunified* theories of self-knowledge. Simple versions of the behaviorist and inner-sense theories are unified: for any mental state M, the account of how one knows one is in M is broadly the same — by observing one's behavior, or by deploying one's "internal scanner." But some philosophers adopt a divide-and-conquer strategy, resulting in more-or-less disunified theories. For instance Davidson (1984) and Moran (2001) offer accounts chiefly of our knowledge of the propositional attitudes, in particular, belief. Knowledge of one's sensations, on the other hand, is taken to require a quite different theory, which neither of them pretends to supply. "[T]he case of sensations," Moran writes, "raises issues for self-knowledge quite different from the case of attitudes of various kinds" (Moran 2001: 10). Similar divisions, although less sharply emphasized, are present in the theories of self-knowledge defended in Goldman 1993 and Nichols and Stich 2003.

There is a case for unification. Suppose that the epistemic capacities one employs to know one's sensations are quite different from the capacities employed to know what one believes. As an illustration, suppose that the account just sketched for belief is correct, and that one knows one's sensations by employing (in part) a dedicated mechanism of inner-sense. Then dissociations are to be expected. In particular, a person's internal scanner could be inoperable, sparing her capacity to find out via BEL what she believes. Such a person would exhibit pain behavior like the rest of us. But does she *feel* an itch in her shin? "Well, probably, since I just caught myself scratching it."

Since such dissociations never seem to occur, this indicates that the epistemology of mental states draws on fundamentally the same capacity, in the sense that individual capacities to know that one is in particular kind of mental state are a package deal. The capacity to know what one believes, for example, brings in its train the capacity to know that one sees a tomato, feels an itch, and wants a beer.

And since the capacity to know what one believes merely involves our general cognitive capacity for reasoning or inference from (typically) worldly premises to mental conclusions (as BELIEF implies), that capacity should also allow one knowledge of one's other mental states. And if it does, then the correct theory of self-knowledge is economical. Roughly put, our general capacity for reasoning about the world suffices for knowledge of our own minds.

But how can the account for belief be extended to desire? According to Nichols and Stich, it clearly can't:

> we can answer… questions about current desires, intentions, and imaginings, questions like: 'What do you want to do?'; 'What are you going to do?'; 'What

> are you imagining?' Our ability to answer these questions suggest that the ascent routine strategy [i.e. the Evans-style procedure[10]] simply cannot accommodate many central cases of self-awareness. There is no plausible way of recasting these questions so that they are questions about the world rather than about one's mental state. As a result, the ascent routine strategy strikes us as clearly inadequate as a general theory of self-awareness. (Nichols and Stich 2003: 194)

Finkelstein agrees:

> [I]t is difficult to claim that the self-ascription of belief [à la Evans] provides a model of self-knowledge that can be used in order to understand our awareness of our own, say, desires because there seems to be no "outward-directed" question that bears the kind of relation to "Do I want X?" that the question "Is it the case that *p*?" bears to "Do I believe that *p*?" (Finkelstein 2003: 161)[11]

If these philosophers are right, and the world-to-mind account that seems so promising for belief is hopeless for desire, then the account for belief should be rejected too. (Nichols and Stich, at least, reject it.) Contrariwise, if the world-to-mind account for belief is right, then there must be a similar account that works for desire. The next section takes up the challenge of finding it.

---

[10] 'Ascent routine strategy' is a phrase of Gordon's (1996), whom Nichols and Stich are specifically criticizing.

[11] See also Goldman 2000: 182-3, Bar-On 2004: 114-8.

4. Desire[12]

In fact, the previous two quotations are far too pessimistic. Although it might superficially appear that "there is no plausible way of recasting" a question about one's desires as a "question about the world," a second glance suggests otherwise. The issue of where to dine arises, say. My accommodating companion asks me whether I want to go to the sushi bar across town or the Indian restaurant across the street.[13] In answering that question, I attend to the advantages and drawbacks of the two options: the tastiness of Japanese and Indian food, the cool Zen aesthetic of the sushi bar compared to the busy garish décor of the Indian restaurant, the bother of getting across town compared to the convenience of walking across the street, and so on. In other words, I weigh the *reasons* for the two options — the "considerations that count in favor," as Scanlon puts it (1998: 17), of going to either place. These reasons are not facts about my present psychological states; indeed, many of them are not psychological facts at all.[14]

Suppose I determine that the Indian option is the best — that there is most reason to go the Indian restaurant. (This might be the result of agonized deliberation; more typically, it will be a snap judgment.) Once I have this result in hand, which is not (at least on the face of it), a fact about my present desires, I then reply that I want to go to the Indian restaurant.

---

[12] For accounts related to the one defended in this section (itself an elaboration of the last few pages of Byrne 2005), see Shoemaker 1988: 47-8, Moran 2001 (especially 114-6), and Fernández 2007. Ashwell 2009 has a critical discussion of Fernández's proposal and the present one.

[13] Although there are differences of usage between 'desire' and 'want,' in this paper the two are treated as equivalent. The semantics and pragmatics of these verbs are relevant to the argument of this paper, but are not discussed for reasons of space.

[14] On reasons as facts see, e.g., Thomson 2008: 127-8.

This example is one in which I "make up my mind" and form a new desire: prior to being asked, I lacked the desire to go the Indian restaurant. But the Evans-style point about looking "outward — upon the world" still holds when I have wanted to go to the Indian restaurant for some time. Of course, often when in such a condition, I can recall that I want to go. But on other occasions the process seems less direct. What immediately comes to mind is the non-psychological fact that the Indian restaurant is the best option; and (apparently) it is by recalling this that I conclude I want to go there.[15]

An initial stab at the relevant rule for desire — specifically, the desire to act in a certain way[16] — is therefore this:

DES* If ϕing is the best option, believe that you want to ϕ

This is not a bad fit for a restricted range of cases, but the general hypothesis that we typically know what we want by following (or trying to follow) DES* has some obvious problems. In particular, the hypothesis both under-generates, failing to account for much knowledge of our desires, and over-generates, predicting judgments that we do not make.

To illustrate under-generation, suppose that I am in the happy condition of also wanting to eat at the sushi bar. Eating at either place would be delightful, although on balance I prefer the Indian option. In such a situation, I can easily know that I want to eat at the sushi bar, despite not judging it to be the best option.[17]

---

[15] Compare the discussion of Moran 2001 in Byrne 2005: 82-5.

[16] Many desires are for other things, of course, some involving oneself and some not: one may want to be awarded the Nobel Prize, or want Pam to get promoted, or want global warming to end, and so forth. These other sorts of desires do not raise any intractable difficulties of their own, and so for simplicity only desires to act in a certain way will be explicitly treated.

[17] As this case illustrates, to want something is not to prefer it over all other options. For reasons of space, this paper concentrates on the epistemology of desire, not the (closely related) epistemology of preference.

To illustrate over-generation, suppose that I really dislike both Japanese and Indian cuisine, and I don't much care for my companion's company either. Still, he would be terribly offended if I bailed out of dinner, and would refuse to publish my poetry. I don't *want* to eat at the Indian restaurant but — as children are often told — sometimes you have to do something you don't want to do. The Indian is the best of a bad bunch of options, and I accordingly choose it. Despite knowing that eating at the Indian restaurant is the best course of action, I do not follow DES* and judge that I want to eat there. Later, in between mouthfuls of unpleasantly spicy curry, I hear my companion droning on about his golf swing, and I think to myself that I really do not want to be doing this.

The description of this example might raise eyebrows, since it is something of a philosophical dogma that intentional action invariably involves desire — on this view, if I slouch to the Indian restaurant, resigned to a miserable evening, I nonetheless must have wanted to go there. Whether this is anything more than dogma can be set aside, because (wearing my Plain Man hat) I will not agree that I want to go to the Indian restaurant. So, even if I do want to go to the Indian restaurant, I am ignorant of this fact, and what primarily needs explaining is the Plain Man's self-knowledge, not the self-knowledge of sophisticated theorists.[18]

---

[18] One of the main contemporary sources for the philosophical dogma is Nagel:

> …whatever may be the motivation for someone's intentional pursuit of a goal, it becomes in virtue of his pursuit ipso facto appropriate to ascribe to him a desire for that goal…Therefore it may be admitted as trivial that, for example, considerations about my future welfare or about the interests of others cannot motivate me to act without a desire being present at the time of action. That I have the appropriate desire simply *follows from* the fact that these considerations motivate me… (Nagel 1970: 29)

In the under-generation example, why do I think I want to go to the sushi bar? Going there is not the *best* option, all things considered, but it is a *good* option, or (much better) a *desirable* one, in the *Oxford English Dictionary* sense of having "the qualities which cause a thing to be desired: Pleasant, delectable, choice, excellent, goodly." Going to the sushi bar is not merely desirable *in some respects*, but desirable *tout court*. The sushi bar is a short cab ride away, the saba is delicious, an agreeable time is assured, and so on. If the Indian restaurant turns out to be closed, that is no cause to investigate other alternatives: going home and heating up some leftovers, getting takeaway pizza, and so on. The sushi bar is a more than adequate alternative. In the over-generation example, by contrast, the Indian option is not desirable, despite being the best.

So these two problems can both apparently be solved simply by replacing 'best' in DES* by 'desirable,' yielding the rule:

DES If ϕing is a desirable option, believe that you want to ϕ.

And the hypothesis corresponding to BELIEF (section 2) is:

DESIRE Knowledge of one's desires is typically obtained by trying to follow DES.[19]

But Nagel gives no actual argument. His conclusion does not follow from the fact that "someone's intentional pursuit of a goal" requires *more than belief*, because there are many candidates other than desire that can take up the slack, for instance intention.

A charitable interpretation is that Nagel is using 'desire' in the technical Davidsonian sense, to mean something like "pro-attitude" (cf. Dancy 2000: 11). That appears to be true of some other philosophers who follow him, such as Smith (1987: ch. 4), although not of Schueler (1995). According to Schueler, Nagel's claim is false in one sense of 'desire,' and true in another "perfectly good sense" of the word (29). However, he provides little reason to think that 'desire' is ambiguous in this way.

[19] Note that DESIRE does *not* imply that there are no other ways of gaining knowledge of one's desires (a similar remark applies to BELIEF, in section 2 above).

If DESIRE is true, then DES is a *good* (knowledge-conducive) rule. So let us now examine whether it is — if it is not, other objections are moot.

The rule BEL, recall, is:

BEL If *p*, believe that you believe that *p*

As noted in section 2, BEL is *self-verifying*: if one follows it one's second-order belief is true. As argued in that section, this observation defuses the objection that following BEL cannot yield knowledge because the fact that *p* is not a reliable indication that one believes that *p*.

A similar objection applies to DES: that ϕing is a desirable option is not a reliable indication that one wants to ϕ. Pam's walking three miles to work tomorrow is desirable, because she'll then avoid hours in an unexpected traffic jam, and get promoted for her foresight and dedication, yet (not knowing these facts) Pam wants only to drive.

Unfortunately, a similar reply does not work: DES is *not* self-verifying. Cases of accidie are compelling examples. Lying on the sofa, wallowing in my own misery, I know that going for a bike ride by the river is a desirable option. The sun is shining, the birds are twittering, the exercise and the scenery will cheer me up; these facts are easy for me to know, and my torpor does not prevent me from knowing them. If I concluded that I want to go cycling, I would be wrong. If I really did want to go, why am I still lying on this sofa? It is not that I have a stronger desire to stay put — I couldn't care less, one way or the other.

Still, this example is atypical. One's desires tend to line up with one's knowledge of the desirability of the options; that is, known desirable options tend to be desired. (Whether this is contingent fact, or a constitutive fact about desire or rationality, can for

present purposes be left unexamined.) What's more, even though there arguably are cases where one knows that ϕing is desirable and mistakenly follows DES, ending up with a false belief about what one wants, the case just described is not one of them. I know that cycling is desirable yet fail to want to go cycling, but I do not follow DES and falsely believe that I want to go cycling. Lying on the sofa, it is perfectly clear to me that I don't want to go cycling. (Just why this is so will be examined later, in section 4.2.)

Thus, although DES is not self-verifying, it is (what we can call) *practically* self-verifying: for the most part, if Pam follows DES, her belief about what she wants will be true. And that is enough to rebut the parallel objection that following DES cannot yield knowledge because the fact that ϕing is a desirable option is not a reliable indication that one wants to ϕ. Again, this does not *entail* that DES is a good rule, but the burden of proof should be on those who think it is not.

As also noted in section 2, BEL is *strongly* self-verifying. That is, if one *tries* to follow it — if one believes that one believes that *p* because one believes that *p* — then one's second-order belief is true. That feature of BEL is the key to explaining privileged access for belief. Similarly, since one's desires tend to line up with one's *beliefs* about the desirability of the options, whether or not those beliefs are actually true, DES is *strongly* practically self-verifying. Privileged access to one's beliefs and desires can be therefore be explained in basically the same way.

At this point a worry about circularity might arise: perhaps, in order to find out that something is desirable, one has to have some prior knowledge of one's desires. If that is right, then at the very least a significant amount of one's knowledge of one's desires remains unexplained. The next section examines some variations on this theme.

4.1. Desirable and desired

According to the *circularity objection*, in order to follow DES, one has to have some knowledge of one's desires beforehand. (The difference between *following* DES and *trying* to follow it only complicates matters while leaving the basic objection intact, so let us focus exclusively on the former.[20])

In its crudest form, the objection is simply that the relevant sense of "desirable option" can only mean *desired* option. If that is so, then DES is certainly a good rule, but only in a trivial limiting sense. Unpacked, it is simply the rule: if you want to ϕ, believe that you want to ϕ. And to say that one follows *this* rule in order to gain knowledge of one's desire to ϕ is to say that one comes to know that one wants to ϕ because one recognizes that one wants to ϕ. True enough, but hardly helpful.

However, this version of the circularity objection is a little *too* crude, leaving no room for any other features to count towards the desirability of an option. (Recall examples of such features quoted from the *OED*: "pleasant, delectable, choice, excellent, goodly.") A slightly less crude version admits that other features are relevant, but insists that a necessary condition for an option's being desirable is that one desire it. Is this at all plausible?

No. As many examples in the recent literature on "reasons" bring out, desires rarely figure as considerations for or against an action, even the restricted set of considerations that bear on whether an action is desirable. The Indian restaurant example is a case in point. Here is another. Suppose I see that an interesting discussion about the

---

[20] A quick way of seeing that the distinction is of no help is just to consider someone who always knows which options are desirable. She always follows DES, and never merely tries to. If the circularity objection applies here, then only desperate measures can save the account elsewhere.

mind-body problem has started in the department lounge, and I am deciding whether to join in and sort out the conceptual confusion. I wonder whether the participants would applaud my incisive remarks, or whether I might commit some terrible fallacy and be overcome with embarrassment, but I do not wonder whether I *want* to join in. Suppose I want to attend a meeting which is starting soon, and that this desire will be frustrated if I stop to join the discussion in the lounge. I do not take *this* to be a consideration in favor of not joining in, but rather (say) the fact that turning up late to the meeting will be thought very rude.

The force of these sorts of examples can be obscured by conflating two senses of 'reason.' Suppose I want to join the discussion, and that is what I do. So *a reason why* I joined in was that I wanted to. Doesn't that show, after all, that my wanting to join in *was* a reason, namely a *reason for* joining in? No, it does not. That I wanted to join the discussion is a reason in the *explanatory* sense, as in 'The failure of the blow-out preventer was the reason why the Deep Water Horizon exploded.' But it does not follow that this fact is a reason in the (operative) *normative* sense, the "consideration in favor" sense of 'reason.'

There is no straightforward connection between an option's being desirable and its actually being desired that would support a version of the circularity objection. Could a connection between desirability and one's *counterfactual* desires do any better?

As an illustration, consider the claim that ϕing is desirable iff if conditions were "ideal," the agent would want to ϕ. All such analyses have well-known problems; for the

sake of the argument let us suppose that this one is correct.[21] (Since the right-hand side is surely not *synonymous* with the left, take the biconditional merely to state a necessary equivalence.) Does this analysis of desirability suggest that sometimes one needs prior knowledge of one's desires to find out that ϕing is desirable?

First, take a case where one is not in ideal conditions. To return to the example at the end of the previous section, suppose I am lying miserably on the sofa. I know that cycling is desirable; I also know, let us grant, the supposed equivalent counterfactual, that if conditions were ideal, I would want to go cycling. In order for circularity to be a worry here, it would have to be established that (a) I know that cycling is desirable by inferring it from the counterfactual, and (b) I need to know something about my present desires in order to know the counterfactual. Now whatever "ideal conditions" are exactly, they are intended to remove the barriers to desiring the desirable — drunkenness, depression, ignorance, and so on. And, although I do not actually want to get on my bike, the enjoyment and invigorating effects of cycling are apparent to me. Regarding (b), it is quite unclear why I need to know anything about my present desires to know that if the scales of listlessness were to fall from my eyes I would desire the manifestly desirable. And regarding (a), the most natural direction of inference is from left to right, rather than vice versa: my knowledge of desirability of cycling — specifically, its enjoyment and invigorating effects — come first, not my knowledge of the counterfactual.

Second, take a case where one is in ideal conditions. I am lying on the sofa, not at all miserable. I know that cycling is desirable, and I know that I want to go cycling. I also

---

[21] For a more sophisticated attempt see Smith 1994: ch. 5. It is worth noting that Smith's conception of an act's being desirable, namely the agent's having "normative reason to do [it]" (132), is broader than the conception in play here.

know, we may grant, that conditions are ideal. Given the equivalence, do I know that cycling is desirable by inferring it from the counterfactual, which I infer in turn from truth of both the antecedent and the consequent? If so, then there is a clear problem of circularity. But how do I know that the antecedent is true, that conditions *are* ideal? Since the chief purchase I have on "ideal conditions" is that they allow me to desire the desirable, the obvious answer is that I know that conditions are ideal because I know that cycling is desirable and that I want to go cycling. But then the epistemological direction is again from left to right, rather than — as the objector would have it — from right to left. If I know that cycling is desirable prior to knowing that conditions are ideal, then (granted the equivalence) I can infer the counterfactual from the fact that cycling is desirable.

The circularity objection is, at the very least, hard to make stick. Let us now turn to some complications.

4.2. DES and defeasibility

To say that we typically follow (or try to follow) rule R is not to say that we always do. The rule WEATHER ('If the skies are dark grey, believe that it will rain soon') is a good enough rule of thumb, but it is *defeasible* — additional evidence (or apparent evidence) can block the inference from the premise about the skies to the conclusion about rain. For example, if one knows (or believes) that the trusted weather forecaster has confidently predicted a dry but overcast day, one might not believe that it will rain soon despite knowing (or believing) that the skies are dark grey.

Given that DES is only *practically* (strongly) self-verifying, one might expect that rule to be defeasible too. And indeed the example of accidie, used earlier to show that DES is only practically self-verifying, also shows that it is defeasible.

In that example, I am lying miserably on the sofa, contemplating the pleasures of a bike ride in the sunshine. This is not just a situation in which I know that cycling is a desirable option but nevertheless do not want to go cycling. It is also a situation in which I do not *believe* that I want to go cycling. Yet if I slavishly followed DES, I would believe that I wanted to go cycling. So why don't I?

I believe that I am not going to go cycling, but that is not why I don't think I want to go: I sometimes take myself to want to ϕ when I believe that I am not going to. For example, I really want to read *Mind and World* this evening, but that is not going to happen because I don't have the book with me.

A better suggestion is that I believe I do not want to go cycling because I believe I *intend* to remain on the sofa. I do *not* believe I intend to avoid reading *Mind and World* this evening, so at least the suggestion does not falsely predict that I will take myself to lack the desire to read *Mind and World*. However, it is obviously not right as it stands. Suppose, to return to the earlier restaurant example, I want to go both to the Indian restaurant and to the sushi bar, and I then form the intention to go to the Indian restaurant, on the grounds that this option is slightly more desirable. When I realize that I have this intention, I will not thereby refuse to ascribe a desire to go the sushi bar: if the Indian restaurant turns out to be closed, I might say to my companion "No worries, I also wanted to eat Japanese."

This highlights a crucial difference between the cycling and restaurant examples: in the cycling case I do not think that remaining on the sofa is a desirable option — I intend to stay there despite realizing that there is little to be said for doing so. I don't think I want to go cycling because, if I did, why on earth don't I go? The means to go cycling are ready to hand, and the alternative is quite undesirable.

In general, then, this is one way in which DES can be defeated. Suppose one knows that ϕing is a desirable option, and considers the question of whether one wants to ϕ. One will not follow DES and conclude one wants to ϕ, if one believes (a) that one intends to ψ, (b) that ψing is incompatible with ϕing, and (c) that ψing is neither desirable nor better overall than ϕing.

That explains why I don't follow DES in the cycling case, and so don't take myself to want to go cycling. Here the action I intend is not the one I think desirable, and neither is it the one I think best, all things considered. More common cases of action without desire are when the intended action *is* taken to be the best, as in the earlier restaurant example with the tedious dinner companion. Dinner at the Indian restaurant will be terribly boring and I won't have a good time; nonetheless, it is the best course of action available, perhaps even beating out other options (like staying at home with a good book) that are actually desirable. I intend to go, but I really don't want to.

Something else needs explaining, though. It is not just that I fail to believe that I want to go cycling — I also know that I lack this desire. I also know that I lack the desire to go to the Indian restaurant. So how do I know that I *don't* want to go cycling, or don't want to go to the Indian restaurant? (Read these with the negation taking wide scope: not wanting to go, as opposed to wanting not to go.)

In the boring dinner example, I know that going to the Indian restaurant is not desirable — indeed, it is positively undesirable. An obvious explanation of how I know that I do not want to go is that I follow this rule:

NODES If ϕing is an undesirable option, believe that you do not want to ϕ.[22]

NODES does not apply in the accidie example, of course, because I know that cycling is desirable. But the earlier discussion of that case already shows how I know that I lack the desire to cycle: if I really have that desire, what is to stop me getting on my bike? The gleaming marvel of Italian engineering is right there, and staying on the sofa has nothing to be said for it.

Conclusion

As the discussion of the last section brings out, the epistemology of desire is not self-contained, in at least two ways.

First, although one's own desires are not among the features that make for the desirability of an option, one's other mental states sometimes are. For instance, I might well conclude that I want to go to the Indian restaurant partly on the basis of the fact that I *like* Indian food: I like, say, andar palak and plain naans. Liking andar palak (in the usual sense in which one likes a kind of food) is not to be equated with wanting to eat it. One may want to eat broccoli for health reasons without liking it; conversely, one may like double bacon cheeseburgers but not want to eat one. Liking andar palak is doubtfully any kind of desire at all. There is no clear circularity worry here, but the considerations of

[22] A similar explanation can be given of the truth of the narrow scope reading — why I also know that I want not to go to the Indian restaurant.

section 3 indicate that the epistemology of likings should be in the same world-to-mind style. And initially, that is not at all implausible: if I sample andar palak for the first time, and someone asks me if I like it, I turn my attention to its flavor. Does it taste good or bad? There is little reason to think that this involves investigating my own mind, as opposed to the andar palak itself: a lowly rat, who is presumably short on self-knowledge, can easily detect good and bad tastes.[23]

Second, the last section suggested that the complete epistemology of desire partly depends on the epistemology of intention. And in any event, given the case for a unified theory of self-knowledge, if intention cannot be squeezed into the world-to-mind format, that casts doubt on the account defended here. At least that is an excuse for another paper.[24]

References

Alston, W. P. 1971. Varieties of privileged access. *American Philosophical Quarterly* 8: 223-41.

Armstrong, D. M. 1968. *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul.

Ashwell, L. 2009. *Desires and Dispositions*. Ph.D. Thesis, MIT.

Ayer, A. J. 1959. Privacy. *Proceedings of the British Academy* 45: 43-65. Page reference to the reprint in Ayer 1963.

———. 1963. *The Concept of a Person and Other Essays*. London: Macmillan.

[23] See, e.g., Berridge and Robinson 2003: 509.

[24] Namely Byrne 2011.

Bar-On, D. 2004. *Speaking My Mind: Expression and Self-Knowledge*. Oxford: Oxford University Press.

Berridge, K., and T. Robinson. 2003. Parsing reward. *Trends in Neurosciences* 26: 507-13.

Boghossian, P. 1997. What the externalist can know a priori. *Proceedings of the Aristotelian Society* 97: 161-75.

Byrne, A. 2005. Introspection. *Philosophical Topics* 33: 79-104.

———. 2011. Transparency, belief, intention. *Proceedings of the Aristotelian Society* Supplementary Volume (forthcoming).

Dancy, J. 2000. *Practical Reality*. Oxford: Oxford University Press.

Davidson, D. 1984. First person authority. *Dialectica* 38: 101-11. Page reference to the reprint in Davidson 2001.

———. 1991. Three varieties of knowledge. *A. J. Ayer Memorial Essays*, ed. A. P. Griffiths. Cambridge: Cambridge University Press. Page reference to the reprint in Davidson 2001.

Davies, M. 1998. Externalism, architecturalism, and epistemic warrant. *Knowing Our Own Minds*, ed. C. Wright, B. Smith and C. Macdonald. Oxford: Oxford University Press.

Evans, G. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Fernández, J. 2005. Self-knowledge, rationality and Moore's paradox. *Philosophy and Phenomenological Research* 71: 533-56.

———. 2007. Desire and self-knowledge. *Australasian Journal of Philosophy* 85: 517-36.

Finkelstein, D. 2003. *Expression and the Inner*. Cambridge, MA: Harvard University Press.

Gallois, A. 1996. *The World Without, the Mind Within: An Essay on First-Person Authority*: Cambridge University Press.

Goldman, A. 1993. The psychology of folk psychology. *Behavioral and Brain Sciences* 16: 15-28.

———. 2000. Folk psychology and mental concepts. *Protosociology* 14: 4-25.

Gordon, R. M. 1995. Simulation without introspection or inference from me to you. *Mental Simulation*, ed. M. Davies and T. Stone. Blackwell.

———. 1996. 'Radical' simulationism. *Theories of Theories of Mind*, ed. P. Carruthers and P. Smith. Cambridge University Press.

Kant, I. 1787/1933. *Critique of Pure Reason*. Translated by N. K. Smith. London: Macmillan.

Locke, J. 1689/1975. *An Essay Concerning Human Understanding*. Oxford: Oxford University Press.

Lycan, W. G. 1987. *Consciousness*. Cambridge, MA: MIT Press.

McFetridge, I. 1990. Explicating 'x knows a priori that p'. *Logical Necessity and Other Essays*. London: Aristotelian Society.

McGinn, C. 1975/6. A posteriori and a priori knowledge. *Proceedings of the Aristotelian Society* 76: 195-208.

McKinsey, M. 1991. Anti-individualism and privileged access. *Analysis* 51: 9-16.

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

Nagel, T. 1970. *The Possibility of Altruism*. Oxford: Oxford University Press.

Nichols, S., and S. Stich. 2003. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Oxford University Press.

Ryle, G. 1949. *The Concept of Mind*. London: Hutchinson. Page references to the Penguin Books edition, 1980.

Scanlon, T. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Schueler, G. F. 1995. *Desire: Its Role in Practical Reason and the Explanation of Action*. Cambridge, MA: MIT Press.

Schwitzgebel, E. 2011. Self-unconsciousness. This volume, chapter 4.

Shoemaker, S. 1988. On knowing one's own mind. *Philosophical Perspectives* 2: 183-209.

———. 1994. Self-knowledge and "inner sense". *Philosophy and Phenomenological Research* 54: 249-314. Page reference to the reprint in Shoemaker 1996.

———. 1996. *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.

Smith, M. 1987. *The Moral Problem*. Oxford: Blackwell.

Thomson, J. J. 2008. *Normativity*. Chicago: Open Court.

Williamson, T. 1995. Is knowing a state of mind? *Mind* 104: 533.

———. 2004. Armchair philosophy, metaphysical modality and counterfactual thinking. *Proceedings of the Aristotelian Society* 105: 1-23.