Self-Knowledge and Transparency
Alex Byrne and Matthew Boyle

I—Alex Byrne

Transparency, Belief, Intention

This paper elaborates and defends a familiar 'transparent' account of knowledge of one's own beliefs, inspired by some remarks of Gareth Evans, and makes a case that the account can be extended to mental states in general, in particular to knowledge of one's intentions.

> How do I know my own mental acts? How do I know what I just decided; how do I know what I believe, what I suspect, what I intend to do? These are one and all silly questions.
>
> —Vendler, *Res Cogitans*

I

Introduction. Knowledge of one's mental states is usually called *self-knowledge*. As is common, the focus here will be on knowledge of one's *present* mental states. For example, knowing that one believes it's raining, that one wants a beer, that one intends to go for a walk, that one feels itchy, that one remembers being at the pub, that one sees a kangaroo. Since knowledge of one's weight or height is not 'self-knowledge' in the intended sense, the phrase has a suspicious Cartesian flavour. It suggests that one's 'self' is a non-physical entity, to be distinguished from a certain human animal, 6′2″ tall and weighing 180 pounds. However, since the terminology is entrenched, it will not be discarded.

A *theory of self-knowledge* is an account of how we have self-knowledge: an account of the means, methods or mechanisms by which one knows that one believes that *p*, and so forth. A useful comparison is with theories of *environmental knowledge*, as we might call it. How does one know so many things about one's immediate physical environment, for instance that this ball is rolling towards that hole? Theories of environmental knowledge range from the highly abstract to the minutely specific. Far over at the abstract

Proceedings of the Aristotelian Society Supplementary Volume lxxxv
doi: 10.1111/j.1467-8349.2011.00203.x

end is the theory that environmental facts affect us through our sense organs, thus enabling us to know them. For details of the more specific end, consult any cognitive science textbook on perception.

Theories of self-knowledge tend towards the abstract. For example, so-called 'perceptual models' of self-knowledge often merely amount to the claim that facts about our mental lives affect us through a dedicated 'internal mechanism', thus enabling us to know them. That is perhaps a start, but leaves far too much unexplained.

Two features of self-knowledge make it of particular interest. The first is that, by and large, beliefs about one's mental states are more likely to amount to knowledge than one's corresponding beliefs about others' mental states—one has *privileged access* to one's mental states. The second is that one has a way of knowing about one's mental states that one cannot use to come to know about the mental states of others—one has *peculiar access* to one's mental states. The two features are independent, in the sense that neither entails the other.[1] But they are connected: the kind of peculiar access that we enjoy presumably explains why we have privileged access. A satisfying theory of self-knowledge will illuminate this connection.

The main point of this paper is that a familiar 'transparent' account of knowledge of one's beliefs can and must be extended to mental states in general. The next two sections introduce and defend the transparency account. §IV criticizes a variant due to Moran, in his *Authority and Estrangement* (2001), and recently defended by Boyle (2009). §V argues briefly for extending the account to perception, sensation and desire. §VI takes up a particularly difficult case —the epistemology of intention.

II

Transparency. The importance of G. E. Moore's much-quoted remarks about the 'sensation of blue' being 'transparent' (1903, p. 446) was pointed out much later by Shoemaker (1990, p. 101).[2] Suppose one sees a blue mug. If one tries to 'introspect' one's experience of the mug, or 'sensation of blue', one apparently comes up empty-handed. The only objects and properties available for aware-

[1] For more on this point, see Byrne (2005, pp. 81–2).

[2] See also Shoemaker (1963, p. 74).

Proceedings of the Aristotelian Society Supplementary Volume LXXXV
doi: 10.1111/j.1467-8349.2011.00203.x

ness are the mug and its (apparent) properties, such as blueness. As Moore says, 'we … see nothing but the blue' (1903, p. 446).

On the face of it, we are not aware of our perceptual experiences or their properties, at least in nothing like the way we are aware of mugs and their properties. Yet, of course, we do know that we have perceptual experiences (at least in a philosophically unloaded sense of 'perceptual experiences'): one might know, for instance, that one sees a blue mug, or that the mug looks blue to one. It is a short step from this to another point that often goes under the name of 'transparency', namely that one knows that one sees a blue mug, or that the mug looks blue to one, *by* attending to the mug. In order to find out that one sees a blue mug, one does not turn one's attention inward to the contents of one's own mind—Moore's remarks suggest either that there is no such procedure or, if there is, it is not necessary. Rather, one turns one's attention outward, to the mug in one's environment. This epistemological observation was first clearly made by Evans (1982, p. 227).[3]

As Evans also noted, a similar point holds for belief. In a familiar passage, he observes that he can answer the question 'Do you think there is going to be a third world war?' by attending to 'precisely the same outward phenomena as I would attend to if I were answering the question "Will there be a third world war?"' (1982, p. 225).

Since perception raises special difficulties of its own, let us concentrate on the more tractable case of belief. What exactly is this 'transparent' procedure for gaining knowledge of one's beliefs? Suppose that I examine the evidence and conclude that there will be a third world war. Now what? Evans does not explicitly address this question, but the natural answer is that the next step involves an *inference from world to mind*: I infer that I believe that there will be a third world war from the single premiss that there will be one.

III

Belief and the Doxastic Schema. Once the 'transparent' procedure is clarified as the inference from the proposition that p to the conclusion that I believe that p, a serious problem is evident: how can such an inference yield knowledge?

[3] Evans himself credits Wittgenstein with a similar insight (Evans 1982, p. 225).

Proceedings of the Aristotelian Society Supplementary Volume LXXXV
doi: 10.1111/j.1467-8349.2011.00203.x

The problem was noted by Gallois (1996). In his terminology, the argument corresponding to the transparent inference is the 'doxastic schema':

$$\frac{p}{\text{I believe that } p}$$

Plainly the doxastic schema is neither deductively valid nor inductively strong. As Gallois says, the transparent inference, or—as we will put it—reasoning in accord with the doxastic schema, 'does not fit any standard pattern of good inference' (1996, p. 47).

This leads to a paradox. On the one hand, the suggestion that one gains knowledge of one's beliefs by reasoning in accord with the doxastic schema is plausible on its face. Evans's remarks about belief, although they contain no actual argument, strike many as one of those things that are obvious once pointed out. On the other hand, the transparent inference could hardly be worse, and so the second-order beliefs it yields will not be knowledge. One of these two claims must go.

Gallois's own response to this is somewhat indirect. He does not try to establish that transparent inferences lead to knowledge. Rather, he argues that unless I reason in accord with the doxastic schema, 'I will form a deeply irrational view of my non-doxastic world. Applying to myself the concept of belief, where that application is not warranted by evidence, allows me to form a more rational picture of the world' (Gallois 1996, p. 76).

Since this conclusion is consistent with the claim that reasoning in accord with the doxastic schema is not knowledge-conducive, Gallois's argument leaves open the apocalyptic possibility that belief without knowledge is the price of rationality. Despite its limitations, the argument is certainly of interest; does it succeed?

Gallois's argument begins with the following case. Suppose—in the terminology of Shoemaker (1988)—that I am 'self-blind' with respect to my beliefs: I am perfectly rational, have the concept of belief, and my opinions about what I believe are formed solely using the third-person methodology I employ to discover what someone else believes. I do not, then, reason in accord with the doxastic schema.

Now suppose 'I start out as a creationist, and end up being converted to the theory of evolution' (Gallois 1996, p. 76). I start out believing that *p* (say, that 'life was created [in] 4004 BC') and end up

Proceedings of the Aristotelian Society Supplementary Volume LXXXV
doi: 10.1111/j.1467-8349.2011.00203.x

believing that not-*p*. Suppose further that I do not have enough evidence for the third-person methodology to reveal either belief. So 'I cannot think of myself as changing my beliefs' (1996, p. 76). The crucial step in the argument is the next, where Gallois argues that this commits me to a 'deeply irrational view of my non-doxastic world':

> Yesterday I could not tell that I held the creationist belief about life on Earth. How, then, will I recollect yesterday's belief about the age of life? Not like this. Yesterday I believed that life was six thousand years old. After all, I have never attributed such a belief to myself. Instead, I will recollect my believing yesterday that life was six thousand years old like this. Yesterday life was six thousand years old. (Gallois 1996, p. 76)

According to Gallois, I believe that life was not created in 4004 BC, and I also believe that yesterday life was created in 4004 BC—more concisely, since 'yesterday' is vacuous, that life was created in 4004 BC. This is not merely to have a metaphysically strange world view, but to have contradictory beliefs. One way—and, we may grant, the only way—of avoiding this result is to reason in accord with the doxastic schema, thus enabling me to 'recollect' that yesterday I believed that life was created in 4004 BC.

Gallois then gives other examples (1996, pp. 75–6), again arguing that if I do not employ the transparent inference in these special cases, then my view is 'deeply irrational'. He then gives a complicated argument for generalizing this result across the board (1996, ch. 5).

The steps in Gallois's overall argument each deserve discussion in their own right, but for present purposes we can just focus on his treatment of the first example.[4] If a rational animal yesterday believes that *p*, and today acquires compelling evidence that not-*p*, it changes its mind. The belief that *p* is lost, and the belief that not-*p* takes its place. Suppose for the moment that the animal in question has no concept of belief. How will it 'recollect yesterday's belief'? *Ex hypothesi*, it will not remember that yesterday it believed that *p*. But neither will it 'recollect' that yesterday, *p*. The animal will simply believe that not-*p*.

Adding the assumption that the animal possesses the concept of belief, and is able to attribute beliefs to itself and others via 'third-person' means, makes no apparent difference. If the animal has no

[4] For more on Gallois's argument, see Brueckner (1999), Gertler (2010, ch. 6).

doi: 10.1111/j.1467-8349.2011.00203.x

evidence that yesterday it believed that *p*, then it will change its mind without realizing that it does so. It will not 'recollect' that yesterday, *p*, but this does not mean that there is anything wrong with its memory. On the contrary, memory would be useless if storing the fact that not-*p* in memory did not prevent one from continuing to 'remember' that *p*. Gallois's argument thus has a false premiss: I will not 'recollect yesterday's belief about the age of life' at all.

The search for a more direct defence of reasoning in accord with the doxastic schema might seem fruitless. The problem can be made more acute by considering knowledge of epistemically worthless beliefs. Suppose, to take Gallois's example (1996, p. 52), I believe that I will meet a beautiful stranger because my fortune-teller says so. My belief (we may suppose) is entirely without justification and (we may further suppose) is false. None of this stops me knowing, in the usual way, that I believe that I will meet a beautiful stranger. But surely an inference from a falsehood—let alone a completely unjustified one!—is not going to result in knowledge. That is, of course, a moral commonly drawn from discussion of Gettier cases.[5]

To see that this difficulty is not insuperable, first note that the doxastic schema is not entirely without epistemic merit. If one reasons in accord with the doxastic schema, and infers that one believes that *p* from the premiss that *p*, then one's second-order belief is *true*, because inference from a premiss entails belief in that premiss.[6] In this sense the doxastic schema is *strongly self-verifying*.[7] In fact, it is not difficult to show that such reasoning typically yields beliefs that are *safe* in the sense that they could not easily have been

[5] See, for example, Harman (1973, p. 47).

[6] See Brueckner (1998, p. 366). A minor qualification: since inference is not instantaneous, there is no cast-iron guarantee that one's belief in the premiss will remain by the time one reaches the conclusion, in which case one's belief that one (now) believes that *p* will be false. Since the chain of reasoning is as short as it gets, this possibility can be ignored.

[7] Why 'strongly' self-verifying? Consider the *gnostic* schema:

$$\frac{p}{\text{I know that } p}$$

Since knowledge is plausibly just as 'transparent' as belief, the gnostic schema stands to knowing that one knows that *p* as the doxastic schema stands to knowing that one believes that *p*.

The gnostic schema is not *strongly* self-verifying, because one may believe that *p* without knowing that *p*. But it is *self-verifying* in the following sense: if one reasons in accord with the gnostic schema, and (as will often be the case) one *knows* the premiss, then one's belief in the conclusion will be true. (Compare Byrne 2011*a*, §4.)

false (Byrne 2005, pp. 96–8).[8]

Second, there is an attractive alternative diagnosis of Gettier cases, namely that they fail to be cases of knowledge because safety is a necessary condition for knowledge, not because no reasoning through false steps is a necessary condition for knowledge. Putting these two points together gives us a reply to the objection that reasoning from the premiss that *p* to the conclusion that one believes that *p* cannot yield knowledge when the premiss is false.

Admittedly, this falls short of a *demonstration* that reasoning in accord with the doxastic schema is knowledge-conducive, because safety is not sufficient for knowledge. However, it is important to keep the dialectical position in mind. Reasoning in accord with the doxastic schema is not being proposed as a novel method of discovering one's own beliefs, but as a plausible description of the main way in which we actually form beliefs about our own beliefs. Unless there is a compelling reason why this method of belief-formation is *not* knowledge-conducive, the working assumption should be that it is. The role of the observations about strong self-verification and safety is merely to allay some objections, not to prove that reasoning in accord with the doxastic schema can yield knowledge.

This 'transparency' account of the epistemology of belief offers a satisfying explanation of both privileged and peculiar access. Privileged access is explained because the doxastic schema is strongly self-verifying. Peculiar access is explained because the method only works in one's own case: inferring that André believes that *p* from the premiss that *p* will often lead one astray.

And finally, the transparency account is an *economical* theory of self-knowledge, appealing only to epistemic capacities and abilities that are needed for knowledge of other subject matters. Here the contrast is with *extravagant* theories—paradigmatically, perceptual models of self-knowledge—which require more resources.

IV

Moran and Boyle on Transparency. Transparency and belief take centre stage in Moran's *Authority and Estrangement* (2001). 'With

[8] The precise characterization of the basic idea of safety is a disputed matter: see, for example, Sosa (1999) and Williamson (2000, ch. 5).

doi: 10.1111/j.1467-8349.2011.00203.x

respect to the attitude of belief', he writes, 'the claim of transparency tells us that the first-person question "Do I believe *P*?" is "transparent" to, answered in the same way as, the outward-directed question as to the truth of *P* itself' (2001, p. 66);[9] a self-ascription of belief that is arrived at by this procedure is said to 'obey the Transparency Condition' (2001, p. 101). Moran is not explicit about how one ends up believing that one believes *P* after one has determined that *P* is true, but for present purposes we may harmlessly pretend that he endorses Gallois's claim that it is by inference.[10]

Moran's basic idea is that the claim of transparency has its 'source in the primacy of a deliberative rather than a theoretical stance towards one's state of mind' (2001, p. 64). Self-knowledge, at least in the central or primary cases, is a matter of 'making up one's mind'.

On the face of it, this looks like a conclusion drawn from an overly restricted diet of examples. *Sometimes* one comes to know that one believes that *p* after considering the question whether *p* for the first time. I may conclude that I believe that Obama will be re-elected after contemplating his successes and failures, his likely opponents in 2012, and so forth. But on many other occasions my mind is already made up, and no deliberation about whether *p* immediately precedes my forming the belief that I believe that *p*. I conclude that I believe that Obama was born in Hawaii, not after considering the evidence, but simply by recalling the fact that Obama was born in Hawaii.[11] The (partial) explanation of why this procedure yields knowledge is

[9] This is not quite correct as it stands: sometimes one's answer to the former is 'no' while one's answer to the latter is 'I don't know'.

[10] This *is* a pretence, since Moran clearly thinks that the Evans-style procedure for discovering what one believes is *not* an inference. He endorses the claim that one knows what one believes 'immediately' (Moran 2001, pp. 10–12), and glosses this in part as 'involv[ing] no inference from anything else' (2001, p. 90). Much hangs on this issue: if the transparency procedure is non-inferential, then this threatens the present project of giving an economical explanation of privileged and peculiar access.

[11] See Byrne (2005, pp. 84–5). A related but quite distinct point is made by Shah and Velleman (2005, pp. 506–7), who observe that if I am trying to find out if I *now* believe that *p*, the last thing I should be doing is opening a *new* investigation as to whether *p*. (Evans's 'third world war' example, mentioned in §III above, can be read—uncharitably but easily—as suggesting this mistaken procedure.) For a response to the Shah and Velleman point as it pertains to Moran's view, see Boyle (2009) and Moran (2011).

To emphasize: the transparency proposal of §III is *not* that one can determine that one believes that *p* at t_1 by examining the evidence at a later time t_2 and concluding that *p*. Nonetheless, Gertler (2011) tries to leverage the Shah-Velleman observation into a series of objections to the transparency proposal (in particular to the presentation in Byrne 2005, which uses a notational variant of Gallois's doxastic schema); further and different objections are given in Moran (2011). Space precludes examination of these objections here.

doi: 10.1111/j.1467-8349.2011.00203.x

exactly the same in both cases: I reason in accord with the doxastic schema, which is strongly self-verifying.

Moran did not address this objection in *Authority and Estrangement*; recently, though, Boyle has provided a defence, arguing that

> Moran's sort of story about how we know our own beliefs must form a fundamental and independent part of any account of self-knowledge, whatever explanation it goes on to give of our capacity to know our own sensations, appetites, and so on. For the sort of self-knowledge Moran describes is a kind that must be available to any self-knower. (Boyle 2009, p. 156)

Boyle's argument has two parts. In the first part he argues that 'if we ascribe to a subject the sort of representational power that can explain comprehending speech, then we at the same time attribute to that subject the kind of power that would allow her—provided she understood the content of the question "Do I believe that *p*?"— to know her belief as to whether *p* by making up her mind' (Boyle 2009, p. 152).

Boyle begins the second part of his argument by saying that the first part only establishes that 'a subject who has the sort of power of representation that can explain comprehending speech must be one who is *entitled* to accompany her sincere assertions with "I believe"' (Boyle 2009, p. 152). The second part of the argument tries to show that a subject who can comprehend speech 'must actually grasp this entitlement'. More specifically, the second part argues that the entitlement must (or could) be grasped by a subject who understands the first-person pronoun. Putting the two parts together, the conclusion is that 'a subject only understands the content of ["I am *F*"] if he understands that the subject of whom he predicates [*F*-ness] is a subject concerning whom he can know certain kinds of facts, not by observing *that they are* so, but by determining them *to be* so' (Boyle 2009, p. 155).[12]

Boyle does not explain how he is understanding 'entitlement', but presumably he intends it in Burge's sense: 'entitlements are epistemic rights or warrants that need not be understood by or even accessible to the subject' (Burge 1993, p. 458). The conclusion of the first part

[12] Although this conclusion is restricted to subjects who understand language, Boyle clearly thinks this ladder can be kicked away, and that the conclusion may be extended to subjects who can entertain the thought that they believe that *p*, even though they may not be language users. This last step of the argument will not be questioned here.

doi: 10.1111/j.1467-8349.2011.00203.x

of the argument, then, may be rewritten thus: if a subject is able to comprehend speech (in particular, understands 'believes' and the like), and believes that *p*, then she has the capacity to know that she believes that *p* by 'making up her mind', where her warrant or justification for this claim is of a kind that need not be 'accessible to the subject'.

For present purposes, we may set aside the second part of the argument, and focus on the first, which (closely following Boyle's presentation) may be set out as follows:

(P1) A comprehending subject must 'be able to reflect on her grounds for holding a given claim true … she must, as it is sometimes put, be able to "play the game of giving and asking for reasons" … she "can say that *p* just when she takes there to be sufficient grounds for supposing *p* to be true' (Boyle 2009, pp. 150–1).

(P2) '[A] subject who can say that *p* just when she takes there to be sufficient grounds for supposing *p* to be true is a subject whose speech already expresses her beliefs: when she (non-deceptively) says "*p*", she will be affirming something she takes to be true' (Boyle 2009, p. 151).

(P3) '[T]o take something to be true just is to believe it' (Boyle 2009, p. 151).

Hence:

(C) '[I]n a way that conforms to Moran's Transparency Condition', a comprehending subject who says '*p*' 'when she takes there to be sufficient grounds for supposing *p* to be true', 'will also be entitled to say "I believe that *p*"' (Boyle 2009, p. 151).[13]

And if we assume that entitlements often lead to knowledge, we may further conclude that such a comprehending subject is often able to *know* that she believes that *p* 'in a way that conforms to Moran's Transparency Condition'—which, recall, we are glossing as 'in accord with Gallois's doxastic schema'.

Although many questions can be raised about (P1), let us grant it,

[13] This conclusion is explicitly epistemological, which seems to be in some tension with Boyle's earlier remark in a footnote that he will not attempt to explain how Moran's account 'secures *knowledge* for the deliberating subject' (Boyle 2009, p. 138 n. 7).

doi: 10.1111/j.1467-8349.2011.00203.x

since the main difficulty is elsewhere. A comprehending subject, then, is one who can assert that *p* when she takes there to be sufficient grounds for supposing (that is, believing) that *p*. Consider a comprehending subject who deliberates about whether Obama will be re-elected, and who takes there to be sufficient grounds for believing he will be. The subject (we may suppose) utters 'Obama will be re-elected'. How does the argument purport to establish that the subject is warranted (specifically, entitled) in also uttering 'I believe that Obama will be re-elected'?

(P2) and (P3) contain the key observation: if a subject sincerely utters ('(non-deceptively) says') 'Obama will be re-elected', then the proposition she asserts is a proposition she believes ('she will be affirming something she takes to be true', and 'to take something to be true just is to believe it'). Hence, if she also utters 'I believe that Obama will be re-elected', her assertion will be true.

We have already met a non-linguistic version of Boyle's observation in §III, namely that the doxastic schema is strongly self-verifying. And, as we saw in that section, that the doxastic schema is strongly self-verifying does *not* show that transparent inferences lead to knowledge; by the same token, it does not show that the subject is entitled to believe the conclusion. Exactly the same points hold for the linguistic case: that our subject's assertion is guaranteed to be true does not show that she knows, or is entitled to believe, what she asserts. So the first part of Boyle's argument does not establish (C).

For all that, (C) is plausibly *true*, at any rate if it is put in terms of knowledge, rather than the somewhat murkier notion of 'entitlement'. §III made the case that transparent inferences lead to knowledge, by pointing out that it is plausible that we make them, and that objections to their knowledge-conduciveness can be deflected. This entails (C), which restricts the transparent inferences to those expressed linguistically by a subject who is able to 'play the game of giving and asking for reasons'.

Although Boyle in places describes the conclusion of the argument in terms of 'making up one's mind' (see the earlier quotation from Boyle 2009, p. 152), the main passage in which the argument is given states the conclusion with this restriction dropped (see the quotations in the statement of (C) above). And clearly the stronger conclusion is warranted if the weaker one is: the key observation in (P2) and (P3) applies equally to the case where one's mind is already

doi: 10.1111/j.1467-8349.2011.00203.x

made up. So why does Boyle think his argument shows that the ‘making up one’s mind’ cases are fundamental? He explains why in the following passage:

> If this is right, then we are in a position to say why the kind of self-knowledge that Moran characterizes is fundamental. It is fundamental because the ability to say what one believes in the way Moran specifies is intimately connected with the kinds of representational abilities that must be possessed by a subject who can make comprehending assertions. (Boyle 2009, p. 151)

Schematically, Boyle is arguing that because the capacity to obtain knowledge in manner *X* follows from the possession of certain basic abilities, *X*-knowledge is an important or fundamental kind of knowledge. But this reasoning contains a gap: what if *X* is a special case of *Y*, a more general way of obtaining knowledge, which also follows from the possession of these abilities? Then it would be (at best) *Y*-knowledge that was the fundamental kind, not *X*-knowledge. And it is precisely this gap that Boyle has not filled. If there is a fundamental kind of self-knowledge in the vicinity, it is knowledge of one’s beliefs obtained by reasoning in accord with the doxastic schema.

V

The Case for the Uniformity Assumption. Since Moran’s account has no purchase on one’s knowledge of one’s mental states which are completely divorced from deliberation, such as sensations and perceptions, it is incompatible with what Boyle calls the ‘Uniformity Assumption’, ‘the demand that a satisfactory account of our self-knowledge should be fundamentally uniform, explaining all cases of “first-person authority” in the same basic way’ (Boyle 2009, p. 141). Accordingly, Boyle draws the corollary that the Uniformity Assumption is incorrect.[14]

There is a certain irony here. As mentioned in §II, Evans himself suggested that the epistemologies of perception and belief are *both* transparent. And although it is certainly harder to explain exactly how such a transparent epistemology of perception works, the intu-

[14] See also Moran (2001, pp. 9–10).

doi: 10.1111/j.1467-8349.2011.00203.x

itive idea is no less attractive than it is in the case of belief. And once the emphasis is taken off the deliberative case, and placed instead on world-to-mind inferences, a uniform economical account of belief and perception becomes a distinct possibility (Byrne 2011*a*).

Evans went even further, suggesting that his account of 'the ways we have of knowing what we *believe* and what we *experience*' provides 'a good model of self-knowledge ... to follow in other cases' (1982, p. 225). So what about some other cases, for instance knowledge of one's own sensations? Transparency has a clear application here too. I know that I *feel* a pain in my elbow, not by attending to myself, or my own mind, but by attending to the painful disturbance *in my elbow*. (My elbow hurts: hence, I feel a pain in my elbow.)

Desire is another example that can arguably be brought under the general world-to-mind model. Do I want to stay home and crochet or go to out to the John Tesh concert? In answering that question, my attention is 'directed outward', in Evans' phrase, to the two options and their merits. (Going to the concert is desirable: hence, I want to go.)[15]

These sketchy remarks indicate that the Uniformity Assumption should be taken seriously, but there is an argument for it that does not depend on enumerating examples where the transparent approach shows promise. If the epistemology of mental states is not broadly uniform, then dissociations are to be expected. One might find, for instance, someone who knows what she believes like the rest of us, but whose independent mechanism for discovering her desires is disabled, leaving her with only a 'third-person' way of knowing what she wants. Such dissociations do not seem to occur, however.

The Uniformity Assumption is consistent with extravagance, but a related failure of dissociation indicates that the correct theory of self-knowledge is also economical. As Shoemaker observes, 'self-blindness' is not an actual condition: there are no individuals who have only third-person access to their mental lives, with spared rational and other epistemic capacities. The obvious explanation of this fact is that rationality and other epistemic capacities are all that is needed for self-knowledge. And since the world-to-mind approach seems to be the only economical theory of self-knowledge that could explain both privileged and peculiar access, this is a reason for taking it to apply across the board.

[15] See Byrne (2005, pp. 99–100; 2011*b*), Fernández (2007); compare Boyle (2009, p. 136).

doi: 10.1111/j.1467-8349.2011.00203.x

Contrariwise, if the epistemology of some mental states cannot be forced into the world-to-mind mould, then the transparency account for belief is on shaky ground. And perhaps the case that appears most obviously recalcitrant is intention. How could one know what one intends by an inference from world to mind? In Moran's terminology, what could the question 'Do you intend to read *Authority and Estrangement*?' possibly be transparent to? Outlining a transparent epistemology of intention is our final task.

VI

Intention.[16] Start with an everyday example. I am deciding whether to go to a dinner party on the weekend, or to stay home and read *Making It Explicit*. I weigh the pros and cons of the two options, and finally plump for the dinner party. That is, I plan, or intend, to go the dinner party. How do I know that I have this intention? The answer hardly leaps to mind, as Anscombe observes:

> [W]hen we remember having meant to do something, what memory reveals as having gone on in our consciousness is a few scanty items at most, which by no means add up to such an intention; or it simply prompts us to use the words 'I meant to …', without even a mental picture of which we judge the words to be an appropriate description. (Anscombe 1957, p. 6)

Going to the dinner party is, I think, the best option out of the two available. So one natural suggestion for the counterpart of the doxastic schema for intention is this:

$$\frac{\phi\text{-ing is the best option}}{\text{I intend to } \phi}$$

One problem for this proposal is illustrated by cases of accidie. Suppose I know that there is everything to be said for going to the dinner party, but I am overcome with listlessness and cannot bring myself to go out. In such a situation, I will have no inclination to reason in accord with the above schema, and believe that I intend to go.

[16] For a proposal similar to the one suggested here (and developed independently), see Setiya (forthcoming). There are important differences as well as similarities; one difference is that Setiya's account is not supposed to involve an inference (see also note 10 above).

doi: 10.1111/j.1467-8349.2011.00203.x

This problem is not necessarily fatal. To say that one may know *P* by inference from *Q* is not to say that the presence or absence of background evidence is irrelevant. We often reason in accord with the following 'precipitation schema':

$$\frac{\text{The skies are dark grey}}{\text{It will rain soon}}$$

But of course this inference is *defeasible*, in the sense that additional evidence (or apparent evidence) can block the inference from the premiss 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.

But what could block the inference in the accidie example? I know I am listless, but sometimes that *doesn't* prevent me from intending to rouse myself; what's more, on such occasions knowing that I am listless does not prevent me from *knowing* that I intend to rouse myself. I also know that I lack the desire to go the dinner party, but sometimes lacking the desire to ϕ does not prevent me from intending to ϕ; what's more, on such occasions knowing that I lack the desire to ϕ does not prevent me from *knowing* that I intend to ϕ.

This problem need not detain us further, because there is a more decisive objection. The suggested schema fails to accommodate cases where no one option is better than all the others. Suppose I am faced with a choice of adding the vodka and then the orange juice, or adding the orange juice and then the vodka. These two options are, I think, equally good, yet I can easily know that I intend to add the vodka *first*.[17] And on the face of it, I know that I have this intention in the way I usually know my intentions.

A more promising idea exploits the close connection between the intention to ϕ and the belief that one will ϕ. As Anscombe (1957) points out, one expresses the intention to ϕ by asserting that one will ϕ. After I have formed the intention to go to the dinner party I might well announce that I will go, thus conveying that I have this intention. Before I form the intention to go the issue is open. Will I go or

[17] Bratman (1985, pp. 219–20) gives this sort of example as an objection to Davidson (1978), which identifies the intention to ψ with the judgement-cum-'pro-attitude' that ψ-ing is 'desirable'. Anscombe (1957, p. 2) suggests and quickly rejects something similar.

Proceedings of the Aristotelian Society Supplementary Volume LXXXV
doi: 10.1111/j.1467-8349.2011.00203.x

not? Forming the intention to go to the dinner party is making up one's mind to go. These and other considerations have led many to claim that intention *entails* belief: if one intends to ϕ, it follows that one believes that one will ϕ.[18] Indeed, Velleman (1989, ch. 4) has argued that intentions simply *are* beliefs of a certain kind.[19] Velleman's view is controversial, but even the mere claim of entailment is disputed. Still, it is not disputed that in paradigm cases like the example of the dinner party, I believe that I will do what I intend.

When deciding whether to go to the dinner party, my attention is directed to the available courses of action, not to my own mind. In the end, I decide to go—'I will go', I say. Do I believe that I will go because I have inferred that I will go from the premiss that I intend to go? If intention *entails* belief, then this is a non-starter: the entailment explains why I believe that I will go, given that I intend to, and no inference from the premiss that I intend to is needed. The suggestion is not much better if the entailment does not hold. Noting my frequent past failures to act in accordance with my intentions, any reasonable observer who knew that I intended to go to the dinner party would not believe that I will go—that would be a rash conclusion to draw from the evidence. Since (we may suppose) I have the same evidence, if the inferential suggestion is correct I should be equally leery about concluding that I will go. Yet I won't be.

That at least allows us to explore the suggestion that things are precisely the other way around: one concludes that one intends to ϕ from the premiss that one will ϕ. That is, perhaps the inference pattern for intention is:

$$\frac{\text{I will } \phi}{\text{I intend to } \phi}$$

which, Gallois-style, can be called the *bouletic* schema.[20]

Like the first suggested schema for intention, there are problems of defeasibility. One can be illustrated with a nice example of Anscombe's: '[I]f I say "I am going to fail this exam" and someone says "Surely you aren't as bad at the subject as that", I may make my meaning clear by explaining that I was expressing an intention, not

[18] See Grice (1971), Harman (1986, ch. 8); for recent scepticism, see Holton (2008).

[19] For a later minor qualification, see Velleman (2000, p. 195 n.55).

[20] 'Boulomai': to will.

doi: 10.1111/j.1467-8349.2011.00203.x

giving an estimate of my chances' (Anscombe 1957, pp. 1–2). I am going to fail this exam, because I skipped the lectures and didn't do the reading. Yet although my failure is evident to me, I do not reason in accord with the bouletic schema and conclude that I intend to fail; to the contrary, I believe that I intend to try to pass.

It might be thought that this problem can be solved with an insertion of 'intentionally':

$$\frac{\text{I will intentionally } \phi}{\text{I intend to } \phi}$$

That appears to cope with the exam example, because although I believe I will fail, I do not believe I will intentionally fail. However, setting aside various issues and unclarities connected with 'intentionally' (which might anyway smuggle intentions back into the picture), this is not sufficiently general, because one can intend to do something *unintentionally.* Altering an example of Davidson's (1971, p. 47), I might intend to trip unintentionally (one means to that end is to walk looking up at the sky), and in such a case I typically have no difficulty in knowing that this is what I intend.

A second problem of defeasibility arises from the phenomenon of foreseen but unintended consequences: one may foresee that one will ϕ, and yet not intend to ϕ, because ϕ-ing is an 'unintended consequence' of something else that one intends. More to present purposes, in such a situation one may *believe* that one will ϕ, and yet not believe that one intends to ϕ. To take an example of Jonathan Bennett's, familiar from the literature on the Doctrine of Double Effect, 'the tactical bomber ... intends to destroy a factory and confidently expects his raid to have the side effect of killing ten thousand civilians' (Bennett 1981, p. 96). The tactical bomber knows he will kill the civilians, yet he does not intend to kill them; in Pentagon-speak, their deaths are 'collateral damage'. If the tactical bomber reasoned in accord with the bouletic schema, he would conclude that he intends to kills the civilians; yet—we may suppose—he disavows having any such intention. Again, to take an example of Bratman's, 'I intend to run the marathon and believe that I will thereby wear down my sneakers' (Bratman 1984, p. 123); I do not intend to wear them down, and neither do I believe that I intend to wear them down.

The examples illustrating these two problems for the bouletic

doi: 10.1111/j.1467-8349.2011.00203.x

schema are cases where one knows what one will do (or, at least, forms a belief about what one will do) on the basis of evidence, but does not ascribe the intention. I know that I am going to fail the exam because I know that I am poorly prepared; I know that I will be wearing down my sneakers because I know that I will be wearing them when I run the marathon. That is, I know on the basis of evidence that I will fail the exam and wear down my sneakers.

However, as Anscombe points out, sometimes one's knowledge of what one will do is not arrived at by these familiar means. (Her official statement exclusively concerns knowledge of what one is doing, but it is clear that the point is supposed to extend to knowledge of what one will do.) As she notoriously puts it, one can know what one is (or will be) doing 'without observation' (Anscombe 1957, p. 13). And those present and future actions that can be known 'without observation' are those that one *intends* to perform: if I know without observation that I will fail the exam, I intend to fail the exam; if I know without observation that I will run in the marathon tomorrow, I intend to run in the marathon tomorrow.

The phrase 'knowledge without observation' is misleading, as it might be expected to work like, say, 'knowledge without googling'. If one knows *P* without googling then, if *P* entails *Q*, one is in a position to know *Q* without googling. But such a principle *isn't* right for 'knowledge without observation', as Anscombe understands it. To adapt one of her examples: I intend to paint the wall yellow, and know that I will paint the wall yellow. That I will paint the wall yellow tomorrow entails that the wall (and paint) will exist tomorrow. But Anscombe does not want to say that I can know that the wall will exist tomorrow 'without observation'.

If we gloss 'knowledge without observation' as 'knowledge not resting on evidence', then this suggests one condition under which the bouletic schema is defeasible, and a first-pass solution to the two problems discussed above. Suppose one knows that one will ϕ, and considers the question of whether one intends to ϕ. One will not reason in accord with the bouletic schema if one believes that one's belief that one will ϕ rests on good evidence that one will ϕ. For example, I believe that I will wear down my sneakers, but I also believe that I believe this because (and only because) I have good evidence for it. Now, from the first-person point of view, an enquiry into one's evidence is (near enough) extensionally equivalent to an enquiry into one's beliefs: one takes *P* to be part of one's evidence

doi: 10.1111/j.1467-8349.2011.00203.x

just in case one believes that one believes *P*.[21] So the defeating condition in effect concerns, *inter alia*, one's own beliefs, the epistemology of which has fortunately already been given an independent account. The complete epistemology of intention thus partly depends on the epistemology of belief.[22]

Like the doxastic schema, the bouletic schema suggests a satisfying explanation of both privileged and peculiar access, although of course the matter needs a deeper treatment. Privileged access is explained because the bouletic schema is *practically* strongly self-verifying: for the most part, if one reasons in accord with the schema (and is mindful of defeating conditions, for instance the one just noted), then one will arrive at a true belief about one's intention. And peculiar access is explained because the method only works in one's own case. First, in the case of others, the defeating condition just noted is almost invariably present: if I believe that André will ϕ, then I will think that this belief rests on good evidence. And, second, the step to the conclusion that André intends to ϕ will in any case often be unwarranted—if André is going to step on an ant, it is most unlikely that he intends to.

Those who find the transparency procedure appealing in the case of belief typically take it to have very limited application elsewhere. Against their position, this paper has argued that the transparency procedure, in Evans' phrase, 'provides a good model of self-knowledge' in general.[23]

Department of Linguistics and Philosophy
Massachusetts Institute of Technology
Cambridge, MA 02139
USA
abyrne@mit.edu

[21] This needs (at least) one qualification, which does not affect the present point. It is possible to believe that one falsely believes *P*; for instance, when one believes not-*P* but has behavioural evidence that one also believes *P*. In that sort of case one will not take *P* to be part of one's evidence.

[22] For those who agree with Williamson's thesis (2000, ch. 9) that one's evidence is one's knowledge, this point is better put by saying that the epistemology of intention partly depends on the epistemology of knowledge (which is just as 'transparent' as belief—see note 7 above).

[23] For helpful discussion and comments, thanks to Caspar Hare, Richard Holton, Julia Markovits and Dick Moran; I am especially indebted to Steve Yablo.

Proceedings of the Aristotelian Society Supplementary Volume LXXXV
doi: 10.1111/j.1467-8349.2011.00203.x

References

Anscombe, G. E. M. 1957: *Intention*. Ithaca, NY: Cornell University Press.

Bennett, Jonathan 1981: 'Morality and Consequences'. *The Tanner Lectures on Human Values*, 2, pp. 45–116.

Boyle, Matthew 2009: 'Two Kinds of Self-knowledge'. *Philosophy and Phenomenological Research*, 78, pp. 133–63.

Bratman, Michael E. 1984: 'Two Faces of Intention'. *Philosophical Review*, 93, pp. 375–405. Page references to the reprint in Bratman 1987.

——1985: 'Davidson's Theory of Intention'. In Bruce Vermazen and Merrill B. Hintikka (eds.), *Essays on Davidson: Actions and Events*. Cambridge, MA: MIT Press. Page references to the reprint in Bratman 1987.

——1987: *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Brueckner, Anthony 1998: 'Moore Inferences'. *Philosophical Quarterly*, 48, pp. 366–9.

——1999: 'Two Recent Approaches to Self-Knowledge'. *Philosophical Perspectives*, 13, pp. 251–71.

Burge, Tyler 1993: 'Content Preservation'. *Philosophical Review*, 102, pp. 457–88.

Byrne, Alex 2005: 'Introspection'. *Philosophical Topics*, 33, pp. 79–104.

——2011*a*: 'Knowing What I See'. In Smithies and Stoljar 2011.

——2011*b*: 'Knowing What I Want'. In JeeLoo Liu and John Perry (eds.), *Consciousness and the Self: New Essays*. Cambridge: Cambridge University Press.

Davidson, Donald 1971: 'Agency'. In Robert Binkley, Richard Bronaugh and Ausonio Marras (eds.), *Agent, Action, and Reason*. Toronto: University of Toronto Press. Page references to the reprint in Davidson 1980, pp. 43–61.

——1978: 'Intending'. In Yirmiahu Yovel (ed.), *Philosophy of History and Action*. Dordrecht: D. Reidel. Reprinted in Davidson 1980, pp. 83–102.

——1980: *Essays on Actions and Events*. Oxford: Oxford University Press.

Evans, Gareth 1982: *The Varieties of Reference*, ed. John McDowell. Oxford: Clarendon Press.

Fernández, Jordi 2007: 'Desire and Self-knowledge'. *Australasian Journal of Philosophy*, 85, pp. 517–36.

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

Gertler, Brie 2010: *Self-Knowledge*. New York: Routledge.

——2011: 'Self-Knowledge and the Transparency of Belief'. In Anthony Hatzimoysis (ed.), *Self-Knowledge*. Oxford: Oxford University Press.

Grice, H. P. 1971: 'Intention and Uncertainty'. *Proceedings of the British Academy*, 5, pp. 263–79.

Proceedings of the Aristotelian Society Supplementary Volume LXXXV
doi: 10.1111/j.1467-8349.2011.00203.x

Harman, Gilbert 1973: *Thought*. Princeton, NJ: Princeton University Press.

——1986: *Change in View*. Cambridge, MA: MIT Press.

Holton, Richard 2008: 'Partial Belief, Partial Intention'. *Mind*, 117, pp. 27–58.

Moore, G. E. 1903: 'The Refutation of Idealism'. *Mind*, 7, pp. 1–30.

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

——2011: 'Self-knowledge, "Transparency", and the Forms of Activity'. In Smithies and Stoljar 2011.

Setiya, Kieran forthcoming: 'Knowledge of Intention'. In Anton Ford, Jennifer Hornsby and Frederick Stoutland (eds.), *Essays on Anscombe's Intention*. Cambridge, MA: Harvard University Press.

Shah, Nishi, and J. David Velleman 2005: 'Doxastic Deliberation'. *Philosophical Review*, 114, pp. 497–534.

Shoemaker, Sydney 1963: *Self-Knowledge and Self-Identity*. Ithaca, NY: Cornell University Press.

——1988: 'On Knowing One's Own Mind'. *Philosophical Perspectives*, 2, pp. 183–209. Reprinted in Shoemaker 1996, pp. 25–49.

——1990: 'Qualities and Qualia: What's in the Mind?'. *Philosophy and Phenomenological Research*, 50, pp. 507–24. Page references to the reprint in Shoemaker 1996, pp. 97–120.

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

Smithies, Declan, and Daniel Stoljar (eds.) 2011: *Introspection and Consciousness*. New York: Oxford University Press.

Sosa, Ernest 1999: 'How to Defeat Opposition to Moore'. *Philosophical Perspectives*, 13, pp. 141–53.

Velleman, J. David 1989: *Practical Reflection*. Princeton, NJ: Princeton University Press.

——2000: *The Possibility of Practical Reason*. New York: Oxford University Press.

Vendler, Zeno 1972: *Res Cogitans: An Essay in Rational Psychology*. Ithaca, NY: Cornell University Press.

Williamson, Timothy 2000: *Knowledge and Its Limits*. Oxford: Oxford University Press.

doi: 10.1111/j.1467-8349.2011.00203.x