Matthew Kennedy University of Nottingham
Matthew.kennedy@nottingham.ac.uk

Naïve Realism, Privileged Access, and Epistemic Safety

Forthcoming in *Noûs*

Working from a naïve-realist perspective, I examine first-person knowledge of one's perceptual experience. I outline a naive-realist theory of how subjects acquire knowledge of the nature of their experiences, and I argue that naive realism is compatible with moderate, substantial forms of first-person privileged access. A more general moral of my paper is that treating "success" states like seeing as genuine mental states does not break up the dynamics that many philosophers expect from the phenomenon of knowledge of the mind.

Although naïve realism is an increasingly prominent theory, its impact in the arena of self-knowledge deserves more examination.[1] In this paper, I outline a naive-realist theory of how subjects acquire knowledge of the nature of their perceptual experiences. From this naive-realist perspective, I also attempt to contribute to ongoing discussion of the relationship between (i) forms of externalism about the mind, and (ii) ideas that contribute to our notion of first-person privileged access.[2]

I will focus on two privileged-access ideas.[3] First, we think that first-person subjects have, with respect to the nature of their own minds, an **epistemic advantage** over third-person subjects. Second, we think that there are **asymmetries between the methods** that first-person parties and third-person parties use to acquire knowledge of the nature of one's mind. I will argue that the naïve-realist conception of veridical experience is compatible with moderate, substantial versions of both of these ideas.

My discussion of first-person epistemic advantage makes use of the concept of epistemic safety. Part of what I say in this discussion may contribute to our understanding of safety's place in the project of epistemically evaluating beliefs. Standard safety principles look for epistemic failure across possible worlds, but they do not measure severity of failure, and this severity can be epistemically important.

As I indicate below, naïve realists describe one's visual experience in a good case as a world-involving success state. A more general moral of my discussion is that, contrary to natural expectations, the treatment of success states as genuine mental states does not destroy first-person privilege.

In the rest of this introduction, I define naive realism, and specify the beliefs that will serve as our working-example beliefs about the nature of one's experience. I also flag a limitation of the discussion to follow.

The naive realist says that in a case of a subject S seeing an object O, S's visual experience (his conscious subjective state) consists in a relation of awareness to O. Since S's experience is a relation to O, O is part of the structure of S's experience. The experience has a world-involving nature in that an external item is part of its structure. The naive realist takes the relational nature of S's experience to have modal consequences. Since O is a part of the structure of S's experience, S could not have that experience unless he were seeing O.

The relational experiential link that naive realism posits between subject and object entails a kind of cognitive success. For the naive realist, the relational, success-entailing features that all philosophers attribute to the condition of *seeing an object* are present within the structure of veridical experience itself. The naive realist says that one's subjective experiential state in a good case is not merely a "wide" state, but a broadly factive or success-entailing state.

Our topic is the first-person access that subjects have to the success-states that the naive realist describes as one's experiences in good cases. In order to pursue this topic, I will examine the properties of first-person belief that one sees a particular object. First-person beliefs of this type attribute the relational, success-entailing features that naive realism attributes to veridical experience itself. For this reason, examining the properties of these beliefs will illuminate naïve realism's impact in the self-knowledge arena. In this examination I will speak of "belief that one sees something" and "first-person knowledge of seeing."

One question that I do not address in this paper is where my account of first-person knowledge of experiential success-states falls with respect to the internalism-

externalism distinction in epistemology.[4] Some resources in my paper (the concept of safe belief, the notion of a disposition to believe) have "epistemic externalist" connotations. On the other hand, the naïve-realist dimension of my proposal may allow for some fit with epistemic-internalist intuitions. It would take some time to establish where my proposal falls on the epistemic internalism-externalism continuum, and it would also take time to assess this placement's effect on plausibility. A prior task is to outline the position that raises these issues, and this is the task I pursue in this paper.

Section I

I will explain first-person knowledge of seeing in terms of subjects' use of perceptual-recognitional abilities.[5] A perceptual-recognitional ability is an ability to tell, by looking, that something is the case. For example, in a reasonably wide range of cases, I can tell by looking whether or not an animal is a chocolate Labrador. My ability to do so reflects my possession of the concept of a chocolate Labrador, and my attunement to the breed's distinctive appearances or looks. My attunement involves a disposition to apply the chocolate-Labrador concept in response to these appearances or looks. Although this disposition occasionally misfires, it produces knowledgeable beliefs most of the time.

The sort of perceptual-recognitional ability indicated here is familiar and widespread. Often we can identify a thing's kind (a dachshund, a golden retriever) on the basis of its look. In these cases, we make a knowledgeable judgment about an item's kind in response to perceptual prompts.

I want to extend this basic structure to self-knowledge. Alongside more familiar judgments like *That's a chocolate Labrador*, I want to explore perceptual-recognitional judgments of the form *I see a chocolate Labrador*. Alan Millar provides us with a good starting point. Millar plausibly argues that many subjects who are able to perceptually recognize brown dogs can also perceptually recognize *that they see brown dogs*.

> …at least in typical situations when we know by looking that something is an F, we are aware, that is know, that we have seen the thing to be an F. In those situations, the visual experiences that, via the exercise of the relevant perceptual recognitional ability, enable us to know that something is an F, also enable us to tell by looking that we see that thing to be an F. This latter knowledge is made possible by the exercise of a higher-order recognitional ability --- an ability to tell whether or not one sees that an F is there (Millar 2008, p. 342).

A conceptually competent subject can do more with the experiences that enable him to recognize F's. On the basis of the experiences that enable him to recognize an F, a conceptually competent subject can also recognize that he sees an F. The same experiences put conceptually competent subjects in position to acquire this additional knowledge.

I have the concept of seeing, and the related (or perhaps ingredient) concept of an object of sight. Moreover, I am perceptually sensitive to instances of the corresponding property, the property of *being an object of sight*. I'm attuned to how objects of sight show up in my own perceptual life. As a part of this attunement, I am disposed, when the issue is salient, to apply the concept of an object of sight to the things that I see. As a result, I can perceptually recognize objects of (my) sight *as such*, as things that I see. Fusing this perceptual-recognitional ability with other perceptual-recognitional abilities, I can make knowledgeable judgments of the form *I see a brown dog* or *I see that brown dog*. Perceptual-recognitional abilities are one resource that I can use to acquire knowledge that I see things. There are other possible resources --- other ways I could acquire this knowledge --- but perceptual-recognitional abilities will be the resource in play in our discussion.

In the next few paragraphs I will try to clarify my proposal about first-person knowledge of seeing by comparing it to some other theories of self-knowledge. At first glance, commitment to recognitional abilities is innocuous. Many philosophers would agree that subjects have the ability to recognize the nature of their experiences. But agreement stated in these terms masks the fact that philosophers cite different resources in

their explanations of how subjects achieve first-person knowledge of their experiences. It's on this more specific level that my recognitional theory is distinctive. For the naive realist, knowledge that one sees something is knowledge of the nature of one's experience. I explain this first-person knowledge of experience in terms of subjects' general perceptual capacities, and in terms of their perceptually sensitive possession of concepts like the concept of seeing and the concept of an object of sight. Other theories of first-person knowledge of experience cite other resources, and thereby take up opposing positions.

For example, an inner-sense theorist explains first-person knowledge of experience in terms of the operation of an inner-sense mechanism (Armstrong 1981, Lycan 1996). Turning to another alternative, Sydney Shoemaker's (1988, 1994) theory of self-knowledge does not appeal to subjects' general perceptual capacities, or to any specific perceptual sensitivities. Shoemaker claims that knowledge of one's mental states is an automatic consequence of one's possession of sufficient amounts of rationality, intelligence, and conceptual competence.[6] If a subject has normal levels of rationality and intelligence, and if he possesses (e.g.) the concepts of belief and experience, then the subject's possession of first-person knowledge of his mental states just follows from these conditions. In this way a subject's self-knowledge supervenes on his rationality, intelligence, and conceptual competence. In addition to our commitment to different resources, Shoemaker and I have different views of what we might call the mechanics of self-knowledge.[7] On my view, a subject's mere possession of the relevant resource does not automatically yield self-knowledge. In order to know that he sees something, a subject must make active use of his perceptual-recognitional ability and make a judgment.

Broadly speaking, I describe first-person knowledge that one sees something as knowledge that is acquired on the basis of one's conceptually informed perception of the "outer" physical world. Another theory that gives knowledge of experience a perceptual

basis is the "displaced perception" theory advocated by Fred Dretske in the mid-nineties.[8] But there are important differences between our theories. After outlining these differences, I will suggest that the divergence leaves my perceptual theory of self-knowledge in a better place.

Dretske writes: "Many of the facts we come to know through ordinary sense perception are facts about objects that we do not perceive" (Dretske 1995, p. 41). By seeing the bathroom scale (when I stand on it), I learn how much I weigh. By seeing the gas gauge on the dashboard, I learn how much gas is left in my car's tank. In cases of displaced perception, subjects learn that *a* is F by seeing *b*, where $b \neq a$, and, often, where *b* is wholly distinct from *a*. The "displaced" object is not perceptually presented to us. Even so, Dretske emphasizes that displaced perception is a form of perception, in that a motorist *sees* that the gas tank is half full by seeing the gas gauge on the dashboard (p. 41).

Displaced perception is guided, and partly structured, by what Dretske calls a *connecting belief*, a belief which asserts a relationship between the perceived object and the object about which we form a belief. Connecting beliefs have something like the following form: *b* would not be G unless *a* were (probably) F.[9]

According to Dretske, we achieve knowledge of our experiences via displaced perception. Suppose S sees that O is blue. Drawing, in addition, on an appropriate connecting belief, S forms the introspective belief *that my experience represents things as blue*. This introspective belief is in effect an abstraction from the apparent character of O, the perceived object. Since O's apparent character --- how O looks --- is a product of content of S's experience of O, if S performs the abstraction correctly he will achieve knowledge of the content of his experience (Dretske 1995, pp. 61-62).

Critics of Dretske have fastened on the presence of connecting beliefs in his account of first-person knowledge of experience. These critics doubt whether subjects

actually have the beliefs that Dretske requires, and they challenge the epistemic standing of these beliefs.[10] In the critique of Drestke, we also find a basic aversion to *inferential* accounts of self-knowledge (Aydede 2003, p. 58). This aversion is not anomalous. Looking beyond the displaced-perception theory, it is widely held that self-knowledge is not acquired by inference (see, e.g., Burge 1988, Heil 1988, Boghossian 1989, Shoemaker 1994). Despite Dretske's protests (1995, pp. 60-62), his commitment to connecting beliefs seems to clash with this picture.

By contrast, recognitional belief is usually described as non-inferential belief. Millar writes

> …recognition is not the drawing of a conclusion from assumptions descriptive of features. In the case of recognising the robin it is a matter of taking in the distinctive *Gestalt* of the bird. Though the relevant features are describable, recognitional knowledge that a bird is a robin is not the upshot of an inference from a set of descriptions. Similarly, the particular *Gestalt* that is shown on an anxious friend's face, is something we take in recognitionally, not by making an inference from assumptions that characterise look and demeanor (Millar 2007b**,** p. 188-189).

Here Millar describes recognitional beliefs as psychologically immediate. We can fill out this description by taking a look at the broad psychological mechanics of recognitional belief.

Someone who has the ability to perceptually recognize chocolate Labradors has a disposition to apply the chocolate-Labrador concept in response to a certain range of his experiences. This disposition is a "good" disposition, in that it tends to produce true judgments. This talk of tending towards the truth indicates that possession of a recognitional ability is in part an environmental and modal condition. Having noted this point about recognitional abilities, I'm going to leave it underdeveloped. [11] In this section I will simply focus on the basic structure of the dispositions that subjects have as part of their possession of recognitional abilities.

When such a disposition is exercised, it produces a belief. The proximate psychological antecedent of the belief is the subject's experience, not another belief. In this way the exercise of recognitional abilities produces non-inferential beliefs.

It will be good to briefly work through some of the preceding points in connection with first-person knowledge of seeing. Judgments of this type seem to have the same phenomenological profile as a recognitional judgment about a perceived animal's kind. Once the topic comes up, it can be immediately clear to me that I see a chocolate Labrador. My account preserves this apparent immediacy by describing first-person judgments about seeing as products of the exercise of belief-forming dispositions whose inputs are one's experiences.

Turning back to Dretske's displaced-perception theory, Dretske and I both try to give first-person knowledge of experience a perceptual basis. Furthermore we agree on the point that this knowledge does not have a *bare* perceptual basis: it does not rest on experience alone. But there is an important difference between our theories. Dretske's additional resource is a specific type of belief. My additional resource is a type of competency, whose psychological inputs are experiences rather than beliefs. For this reason my theory is a better match with the idea that self-knowledge is non-inferential.

Section II

We think that first-person knowledge of one's mind is better or more secure than third-person knowledge of another's mind. There are many ways of expounding the idea that first-person belief has better epistemic standing than third-person belief. We might claim that first-person beliefs about one's mind are infallible, we might claim that mental properties are self-intimating, in that their instantiation necessarily produces first-person belief in their presence. Although our tradition makes these claims familiar to us, they don't have default standing in current discussions of self-knowledge, and I will make no attempt to validate them. Here I will argue that the naive-realist theory of veridical

experience is compatible with a moderate, but nonetheless substantial, conception of the epistemic advantage of the first-person.

We can arrive at my target conception by taking some recent ideas from Alex Byrne, and making some important modifications. Byrne's initial statement of first-person epistemic advantage is this:

> Roughly: beliefs about one's mental states acquired through the usual route are more likely to amount to knowledge than beliefs about others' mental states (and, more generally, beliefs about one's environment) (Byrne 2005 , pp. 2-3).[12]

I adopt the idea that first-person beliefs are more likely to amount to knowledge than other types of belief. I pursue a variant of this idea in coming sections. What I do not adopt is the idea that *environmental beliefs* are less privileged than beliefs about one's mental states. The current-interest beliefs about one's mental states overlap in content with beliefs about the nature of one's environment. A belief that one sees a dog is in part a belief about the current makeup of one's environment. Beliefs about seeing and beliefs about one's environment have some of the same commitments. For this reason, the naive realist cannot deliver on the idea that one's beliefs about one's mind are in better epistemic shape than one's beliefs about one's environment. Nonetheless, there is an important epistemic-advantage issue that remains open. This is whether first-person beliefs about one's mind are in better epistemic shape than third-person beliefs about the minds of others. This is the issue I examine in the next few sections.

Shortly after the preceding quotation, Byrne makes a more expansive statement of his conception of epistemic advantage. He writes

> ...although error may always be a possibility, in a typical situation it is easier to be right about one's (non-factive, non-object entailing) mental states (that one believes the cat is indoors, say) than about the mental states of another (that Fred believes that the cat is indoors), or [*environmental example omitted*] (Byrne 2005, p. 3).

Byrne's claim about the epistemic advantage of first-person belief over third-person belief is conditional on a restriction to *typical situations*. A very important point about my

upcoming discussion is that it incorporates Byrne's restriction to typical situations. This restriction excludes a wide range of atypical situations. In the latter situations, the epistemic advantage of first-person subjects may shrink or vanish entirely (for some examples, see section VI). Byrne's restricted-domain epistemic-advantage thesis is compatible with the existence of cases where a third-person party knows my mind as well or better than I do. Byrne does not try to defend the idea that every possible subject has, on the topic of his own mind, an epistemic advantage over every possible third-person subject. Byrne's idea is that if we restrict ourselves to typical subjects and situations, belief about one's own mind is epistemically more solid than belief about someone else's mind.[13] More precisely, typical first-person belief is "significantly" more solid than typical third-person belief (Byrne 2005, p. 27).[14]

This is the conception of epistemic advantage that I will try to deliver in this paper.

Another important point about my discussion is that I think that Byrne does not need to exclude object-entailing states from his list of the mental states to which we have privileged access. Typical first-person subjects have an epistemic advantage with respect to the object-entailing mental state of *seeing an object in one's environment.* I will try to show that the epistemic advantage of the first person extends to at least one object-entailing state.

Once we make the restriction to typical subjects, we should focus on cases in which a typical first-person subject S ascribes a mental property M to himself, and a typical third-person party J ascribes this same mental property to S. "Same property" cases give us the sharpest test of the idea that the first-person perspective itself delivers epistemic advantage.

A same-property case involving seeing takes the following form:

(i) I believe *I see a brown dog*.

(ii) John believes *Patrick sees a brown dog*.

I will call my belief-state "(S)" and John's belief-state "(J)." In the last section I suggested that S-type beliefs could be instances of perceptual-recognitional knowledge. Here I'm going to work on the assumption that J-type beliefs can be examples of perceptual-recognitional knowledge too. You can perceptually recognize that someone else sees something. You are in your car, waiting for the light at the crosswalk. While you are waiting, you see a pedestrian who is, for the moment, oblivious to his legally sanctioned opportunity to cross the street. His inattention catches your attention, and *then*, he notices the "walk" sign and crosses the street. As you keep your eye on the pedestrian, at the key moment, you recognize that he sees the "walk" sign.

In my comparison of first-person belief about seeing and third-person belief about seeing, I'll make use of the epistemic concept of safety.[15] One appealing feature of this concept is that it enables us to make fairly fine-grained epistemological judgments. We need this ability in order to make headway on the topic of first-person epistemic advantage. I can know that I see a dog, and you can know that I see a dog too. If first-person subjects have an epistemic advantage, we must characterize it in terms of concepts besides knowledge itself, and my concept of choice is safety.[16] In turning to safety, I don't need to assume that safety is necessary for knowledge. All I need is the more general assumption that the concept of safety gives us a useful measure of a subject's epistemic position.

Turning then to this concept, Gettier cases seem to show that subjects who have knowledge are not simply lucky to be right. These subjects' beliefs have some insulation from error. "Safety" theory construes this insulation in modal terms. First try:

> (Safety-1) S's belief that p is safe iff S could not easily have falsely believed that p.

"Ease" is analyzed in terms of the notion of close possible worlds. This notion falls out of the familiar idea that possible worlds can be ordered according to their similarity in relevant respects to the actual world. Closer worlds are more similar worlds. According to Safety-1, S's belief that p is safe just in case there is no close world in which S falsely believes that p**.**

When we ask whether a belief is safe, we ask whether its counterpart doxastic failures fall outside a threshold of closeness such that all operatively close worlds are at least that close.[17] We can also ask whether one belief is *safer* than another.[18] Here we ask whether belief B1's counterpart doxastic failures are more distant than belief B2's counterpart doxastic failures. In this paper I am more interested in the second type of safety question, as applied to first-person and third-person beliefs about seeing. In the rest of this section, I develop a specific notion of counterpart doxastic failure, which in turn will yield a more precise characterization of "safer belief."

Although safety will be a fruitful concept for our queries, Safety-1's "false belief that p" is too blunt an instrument for the purposes of epistemic evaluation. First, some nearby cases of false belief that p are not relevant to the epistemic status of a subject's actual-world belief that p.[19] Suppose I read in the newspaper, and thereby believe, that the Boston Celtics beat the Los Angeles Lakers over the weekend. But suppose that I am also somewhat prone to forming beliefs about basketball outcomes on the basis of coin-flipping. Then, we suppose, there is a close world in which I flip a coin and thereby believe that Boston won, when they did not. (Boston and L.A. are both strong teams). In this close world, my belief is false. But this counterfactual failure should not destroy the epistemic merits of my actual belief, which was formed by a relevantly different method. The specifics of belief formation are epistemically important, and our modal evaluations of beliefs should be sensitive to this point. Safety principles should include a "methods" provision. In our assessments of the epistemic standing of actual-world beliefs, we should

only consider close-world counterfactual beliefs that are formed by the same or similar methods.

Second, safety principles should admit cases in which non-identical but relevantly similar belief-forming methods are used. Our topics in this paper make a pressing case for this sort of tolerance. According to the naïve realist, when a subject visually recognizes that he sees a dog, he employs an object-involving belief forming method, a method that is partly constituted by the subject's experiential relation to the dog. But when the subject has a "matching" hallucination as of a dog, and forms a similar belief, the object-involving method from the seeing case is by definition not available. The naïve realist says that the subject employs a different method in the hallucinatory case. However, if the hallucination occurs in a close world, and causes the subject to falsely believe that he sees a dog, this mistake is epistemically relevant. The modal proximity of this hallucination indicates that the subject, in the actual-world case of seeing, is lucky to be right.

We can grant the naïve realist his conception of methods, while preserving the epistemic relevance of hallucinations that are close *and* matching, by construing the safety principle in a way that treats relevantly similar methods as admissible.

> (Safety-2) S's belief that p is safe iff S could not easily have believed p falsely by employing a relevantly similar method.

Although Safety-2 is a big improvement on Safety-1, there is one remaining difficulty, which can be brought out by taking a closer look at beliefs formed on the basis of hallucinations. Go back first to our example from earlier in this section:

(i) I believe *I see a brown dog*. (= "belief S")

(ii) John believes *Patrick sees a brown dog*. (= "belief J")

I want to supplement our example in a way that acknowledges the following point: in a typical one-dog case, John and I will not be neutral on which dog I see. Our respective outlooks will include commitments on this issue. Each of us will have, or at least aim to

have, a particular dog in mind. In order to capture this feature of our psychologies, I will update my portrayal of the respective beliefs to explicitly include demonstrative components.

(i) I believe *I see a brown dog (that dog)*.

(ii) John believes *Patrick sees a brown dog (that dog).*

Reflecting the update, let's call the first belief (S*) and the second (J*). Take (S*) and consider a world in which I have a full-blown matching hallucination as of a dog. (A "full-blown" hallucination is one in which the subject has no perceptual contact with his immediate environment. See section III, "Type D" for discussion of partial hallucinations). Suppose that I aim to form a belief like (S*) in reaction to my hallucinatory experience. The belief I form includes a failed demonstrative component, and so the belief differs in content from an isomorphic belief in an actual-world case of seeing. (The claim that the beliefs differ in content is all I need for present purposes. We don't need to adopt a view on exactly how the contents differ). Say that my actual-world (S*)-belief is the belief that p. The full-blown hallucination world is not a world in which I falsely believe that p. My cognitive situation has eroded to a point where I do not even have this belief. But the full-blown hallucination world, if close, is epistemically relevant. What this means is that some counterfactual beliefs with different content are epistemically relevant to actual-world beliefs that p.

David Manley (2007) addresses this point and others with the notion of a "failed counterpart thought."[20] Although Manley's notion is broader, for our purposes we can work with a very specific notion of failed counterpart thoughts. Our notion will incorporate the beliefs (S*) and (J*). Failed counterpart thoughts are (i) counterfactual beliefs that are produced by the same methods that produce (S*) or (J*), or produced by methods relevantly similar to these methods, and (ii) are either false or contain failed demonstrative elements of the sort produced in hallucination.

Our working safety principle, which again is defined partly in terms of local concerns, is as follows:

> (Safety-3) S's belief that p is safe iff S could not easily have had a failed counterpart thought.

We can use the concept of safety in our assessment of naïve realism's impact on the epistemic advantage of the first person. We can ask, are first-person beliefs about seeing safer than third-person beliefs about seeing? A belief is safer the more distant are the closest worlds in which its failed counterparts occur. A belief B1 is safer than another belief B2 just in case B1's failed counterparts are more distant than B2's.

Section III:

Here are the protagonists in our comparison of first-person and third-person beliefs.

> (i) I believe *I see a brown dog (that dog)*. [= (belief S*)]
>
> (ii) John believes *Patrick sees a brown dog (that dog).* [= (belief J*)]

In what follows I will speak of beliefs (S*) and (J*) as if they are actual-world token beliefs. But my intent is for (S*) and (J*) to represent the far more general phenomena of typical first-person belief about seeing, and typical third-person belief about seeing. I will assume that John and I are typical subjects. Part of what I have in mind here is that both John and I possess the concept of seeing, and that we both competently apply this concept and other relevant concepts on the basis of our experiences. I will leave the "typical" restriction implicit in much of what follows.[21]

I will defend something close to the following claim, which I provide as a preview for orientation purposes.

> (MUCH SAFER) First-person beliefs about seeing are much safer than third-person beliefs about seeing.

Here are three scenarios in which (S*) and (J*) are held and come out false.

> TYPE A: I don't see anything.
> TYPE B: I see something, but it's not a dog.
> TYPE C: I see a dog, but it's not brown.

These scenarios introduce three types of error, three ways in which John and I could believe (S*) and (J*) and be wrong. I will proceed by making some judgments on how safe (S*) and (J*) seem to be from these types of error, and from other types of error that will come up in the course of our discussion. Although such judgments are not of paradigm solidity, I believe that my judgments will strike enough chords to make a good case for the epistemic advantage of the first-person.

TYPE A, FIRST PERSON FALSE BELIEF.

Here I believe, on experiential grounds, that I see a brown dog, but in fact I don't see anything. How could this happen? It could happen if I hallucinated as of a brown dog. More expansively, it could happen if I had a full-blown, perfectly life-like hallucination as of a brown dog. It's not immediately clear how else it could happen. No other recipe readily suggests itself. It seems we have to turn to so-called "matching" hallucinations in order to explain my experientially grounded, but false belief.

But a world in which I have a full blown matching hallucination as of a brown dog, a hallucination that would cause me to non-inferentially believe that I see a brown dog, seems to be very distant. For this reason, my actual-world belief that I see a brown dog is very well insulated from Type A failure. Thus, with respect to Type A failure, my belief (S*) is very safe.

TYPE A, THIRD PERSON FALSE BELIEF.

In this sort of case, John believes, on experiential grounds, that I see a brown dog, but in fact I don't see anything. Cases of this sort are of course very diverse. But there is at least one mundane circumstance that could trigger a Type A error for John. Suppose my eyes appear to be open at the key moment when in fact they are not open, or at least they are not open enough for me to see anything.

The third-person "eyes-closed" possibility need not be very close. We can grant that John's belief (J*) is safe from Type A error. Even so, the eyes-closed scenario seems

to be a closer possibility than the first-person full-blown hallucination scenario. In fact the former possibility strikes me as much closer. If "eyes-closed" is much closer, John is much less safe. The perceptual-recognitional judgments of *first-person* subjects are much safer with respect to Type A error. See below for more discussion of this point.

TYPE B: I see something, but it's not a dog.

TYPE B, FIRST PERSON FALSE BELIEF.

In this case, I believe, on experiential grounds, that I see a brown dog. I do see something, but it's not a dog. Perhaps I see an eligible looking animal, but only from far away, and from this distance I erroneously judge it to be a dog.

This is a natural-sounding scenario, and it's the one I will consider for present purposes. Misjudging an animal's kind from long distance is a mundane mistake. Type B error worlds are non-exotic, even for a competent subject. Even so, we should not regard Type B mistakes as especially close possibilities. The closest scenarios in which a *competent* subject views an animal from an objectively difficult distance should not be scenarios in which the subject judges falsely, but rather scenarios in which he either gets it right or withholds belief on the animal's kind. Type B error worlds are a bit further out. If I'm a relevantly competent subject, my belief (S*) is fairly safe from Type B error.

With this point in hand, I want to briefly return to the third-person "eyes-closed" case from Type A, which I believe can be described in similar terms. Since John is a competent subject, he does not normally make Type A errors. John tends to get it right or withhold belief in difficult cases. Nonetheless, Type A "eyes-closed" error is non-exotic for him, and thus for third-person subjects generally. By contrast, hallucination-induced Type A First Person error is very exotic, at least for typical subjects. Therefore first-person beliefs about seeing are much safer from Type A error.

Let's go back to Type B.

TYPE B, THIRD PERSON FALSE BELIEF.

John believes, on experiential grounds, that I see a brown dog. Here, I do see something, but it's not a dog. On one reading of Type B, Third Person, it is not materially different from Type B, First Person. Here, John misjudges the kind of the object (call this object O) that I see. He gets everything else right: I do see O, it's just that John is wrong about O's kind. John's mistake in this case is structurally identical to my mistake in Type B, First Person. I assume that John is equally well insulated from this sort of error. My belief S* is not any safer than John's from this sort of error.

But we should also consider another way for John to mess up. Consider a case in which I see an object, but I do not see the object that John thinks I see. For example, suppose that John and I are standing in different parts of a lightly populated but somewhat variegated plaza. John thinks I see a certain brown dog that he picks out on the basis of his experiences. John's belief about what I see is demonstrative --- *Patrick sees that dog*. The reference of his mental demonstrative is determined by his perceptual attention to the dog in question. But in fact my perspective at that moment does not allow me to see this dog. Instead I see a cyclist who temporarily occludes the dog.

This "cyclist" case meets the bare letter of TYPE B, THIRD PERSON. But what's going on deserves its own category. Before we turn away from Type B we should remind ourselves that the beliefs S* and J* are equally safe from the Type B errors that I described in earlier paragraphs.

The cyclist case is better characterized as follows:

> **TYPE D: I see something, but I do not see the object that is perceptually picked out and believed (by the relevant subject) to be seen by me.**

Although I have introduced Type D in the course of discussing the third-person, my discussion of Type D will start with the first-person case and then return to the third-person case.

TYPE D, FIRST PERSON.

The "relevant" subject indicated in Type D is the first-person or third-person party whose beliefs we are evaluating. Type D third-person cases are easy to construct: the third-person subject perceptually picks out an object and believes (demonstratively) that I see it, but I don't see this object. Instead, I see another object.

Type D is more difficult to implement for the first person. Here, I perceptually pick out an object O, and I believe that I see it. My belief is demonstrative in character --- I believe *I see that dog* --- and its reference is anchored by my perceptual attention to O. But I do not see O. I see something else.

But this is, to say the least, hard to model: it seems that if I visually discriminate and identify an object O, then I see O.

If Type D error is impossible for first-person subjects, and possible for third-person subjects, this would be a welcome result for defenders of first-person privilege. However, holding off on that outcome, if we agree to talk liberally about objects of experiential attention, and about objects of experiential demonstrative belief, while remaining strict about the concept of ***seeing***, we can find a model for Type D first-person error.

Consider partial hallucinations. The "full-blown" hallucinations of earlier examples involve no mental contact at all with one's immediate environment. Partial hallucinations are somewhat more anchored in reality. Suppose that I hallucinate as of a brown dog, while making perceptual contact with parts of my immediate environment. The "as of" locution is ontologically neutral. For present purposes, I'd like to stay on an ontologically neutral level while introducing slightly different terminology. Another way to describe the just-introduced partial hallucination is as follows: an ersatz dog appears to me against a non-ersatz indoor background. I experientially attend to this ersatz dog, and it thereby extracts from me a failed counterpart of belief S*.[22] In this scenario I do see

things. But I do not see the object to which I experientially attend, I do not see the object of my failed counterpart thought.

I assume that since partial hallucinations are less intense than full-blown hallucinations, they occur in closer worlds. Still, the canvassed partial hallucination as of a brown dog is not a close possibility at all. Partial matching hallucinations are still quite exotic. I am very safe from Type D error.

TYPE D, THIRD PERSON.

However, John is not very safe from Type D error. He messes up in Type D fashion when I do not see the object that John perceptually picks out and thinks I see. John has false commitments in a Type D case; he has an inaccurate conception of my subjective situation.

John could make an error with respect to my vantage point, misjudging its lines of sight. Or John could accurately assess what *ought* to be perceptually available to from my vantage point, only to be thwarted by an interloper whose influence is unknown to John. The temporarily occluding cyclist is one example of the latter possibility.

As before, we need not regard these possibilities as very close possibilities. We can stipulate that John is safe from Type D error. As a statement about the general phenomenon of third-person judgments about seeing, this may be overgenerous. We should not underestimate the linking-up task that third-person subjects face, and we should not overlook the significant epistemic disparity that this task imposes. Even if John meets a safety threshold, the problem is that he is not safe enough. Type D third-person error possibilities are entirely mundane as compared to the still quite exotic prospect of matching, but only partial, hallucinations. I have a big epistemic advantage on John when it comes to Type D error. My advantage is almost as big as the advantage that obtains with respect to Type A error. My error possibilities are quite exotic; John's are not exotic at all. I'm a lot safer than he is from Type D error.

Let's briefly touch base with Type C. I will restate it and Type B so that they clearly mark out types of failure different from Type D.

> **TYPE B*: I see an object (and the evaluation-relevant S*/J* belief apprehends that I see this object), but the object is not a dog.**
>
> **TYPE C*: I see a dog (and the evaluation-relevant S*/J* belief apprehends that I see this dog), but the dog is not brown.**

On inspection, Type C does not introduce any new issues. Once we distinguish Types B and C from Type D, C is a replay of B. Going back to Type B*, there are non-exotic scenarios in which I falsely believe that an animal I perceive is a dog, because I view the animal from long distance. The same is true of John.

Turning to Type C*, there are realistic scenarios in which I falsely believe that a dog I see is brown, perhaps due to poor lighting. Since I'm a doxastically responsible subject, these scenarios are not especially close, but they are not exotic either. We should assume, again, that John is in the same boat: that John is equally prone to Type C* error as induced by bad lighting.

Section IV

Remember that a belief is safer the more distant are the closest worlds in which its failed counterparts occur. A belief B1 is safer than belief B2 just in case B1's closest failed counterparts are more distant than B2's closest failed counterparts. The scorecard from section III is that John and I are equally safe from B* and C* errors, and I am much safer than John from A and D errors. In order to draw an overall epistemic moral from this scorecard, we need to identify the closest failed counterparts of my belief S* and John's belief J*, and to see which counterparts are closer. This will show whether and to what extent first-person beliefs about seeing are more secure than third-person beliefs.

Although I do not think we have settled on exactly the right procedure, the results of the current one are worth getting into view. Since I am very safe from A/D error, my belief S*'s closest failed counterparts are of the B*/C* type. Since John and I are on a par

with respect to safety from B* and C* error, interest turns to third-person A/D errors. We must ask whether **third-person A/D errors** are closer than **B*/C* errors**. Any difference in proximity between the closest failed counterparts of S* and J* will be between these categories.

However, the rough judgment from the last section was that third-person A/D errors are not closer than B*/C* errors from either subject. I suggested that all of the mistakes just mentioned are not especially close, but are not exotic either. Third-person A/D errors share vicinity with B*/C* errors, including first-person B*/C* errors. I see no payoff in trying to refine this judgment. Even if some slim margin emerged favouring either the first-person or the third-person, any resulting epistemic advantage claim would be too narrow to be worth defending. The closest failed counterparts of S* and J* are equally close.

Thus the apparent conclusion from our comparison of S* and J* is that first-person beliefs about seeing have no epistemic advantage over third-person beliefs about seeing, since these beliefs are equally safe. Apparently, naïve realism's externalist conception of experience undermines the epistemic advantage of the first person, at least in connection with knowledge of experience.

But I think this is the wrong lesson to take from our discussion of the modal proximity of A, B*, C*, and D-type errors. Part of the problem lies with the conceptual apparatus of safety. In its current form, this apparatus leaves out an important dimension of the comparative shakiness of actual-world beliefs. And part of the problem lies in the way I have been funnelling types of doxastic failure into our safety machinery. My policy so far has overlooked important relationships between categories A-D, and overlooked the different ways in which these categories connect with our current interest in self-knowledge.

Once we get clear on these points, we can defend more naïve-friendly verdicts on the epistemic advantage of the first-person.

To start with, consider the evolution of categories B and C over the course of discussion in the last section. We started with

> **TYPE B: I see something, but it's not a dog.**
> **TYPE C: I see a dog, but it's not brown.**

After introducing Type D, we revised Types B and C to distinguish them from the former sort of error. The revised categories were:

> **TYPE B*: I see an object (and the evaluation-relevant S*/J* belief apprehends that I see this object), but the object is not a dog.**
>
> **TYPE C*: I see a dog (and the evaluation-relevant S*/J* belief apprehends that I see this dog), but the dog is not brown.**

(A terminological reminder is that "Type B* failure," for example, occurs when John or I have counterfactual beliefs isomorphic to beliefs S* and J*, and these counterfactual beliefs come to grief in the conditions described in TYPE B*).

These revisions of B and C indicate the following connection between these categories and the other categories A and D: subjects who fail at the B*/C* levels are subjects who *succeed* at the levels of A and D. (A/D success is just avoidance of A/D error). Our revised categories make clear that belief-acts which make B* or C* mistakes are belief-acts which avoid error at the levels of A and D. For this reason, B*/C* failure involves cognitive success at the levels of A and D.

As a result, attributions of B*/C* error to subjects are not unequivocal attributions of failure. Rather, attributions of B*/C* error encode significant forms of cognitive success. Specifically, subjects who make B*/C* mistakes actually get a good amount right about my mental life. These subjects accurately apprehend that I see something, and they accurately apprehend, as such, the object that I see.

These forms of success are important in our context, given our interest in the phenomenon of knowing one's mind. We can recognize this importance by assigning a

prominent place, in our evaluations of counterparts of S* and J*, to the project of accurately apprehending my subjective orientation to items in the world. B*/C* *errors* do well with respect to this project. Of course, subjects who make B* and C* errors have false beliefs. But these subjects do not fail severely at the project of accurately apprehending my subjective perspective on the world. As assessed in terms of this project, B*/C* mistakes are not *big mistakes*.

By contrast, measured in terms of our prioritized project, A/D mistakes are big mistakes. An agent who makes a Type A error misses the fact that I'm not seeing anything; and an agent who makes a Type D error does not accurately identify the object that I see. These agents get important things wrong about my subjective orientation to items in the world. Thus they make big mistakes in a way that B*/C* subjects do not.

Once we are interested in the severity of particular types of error, we should notice that safety principles like Safety-3 do not measure this quantity. Safety-3 charts the modal proximity of failed counterparts, but it does not measure the severity of their failure. However, on the face of it, the modal proximity of severe failures is epistemically interesting.[23] Note that the notion of a severe failure is project-relative and thus context sensitive. Even so, within a context, it is natural to ask about the proximity of the errors that the context identifies as big mistakes. How safe is your belief that p from a big mistake? How easily could you have washed out (more or less) on the target project?

The severity of a belief's failed counterparts seems to have an impact on our judgments about the belief's comparative epistemic status. Suppose that two people believe that p in a way that their closest failed counterpart beliefs are equally close. If context says that one of these counterparts is a big mistake and the other is not, then the person with less extreme error on his modal scorecard seems to be in better epistemic shape. His grasp of the contextually-prized issue is firmer because his mistakes within the specified modal sphere are less severe.

If we add severity of failure to our calculations, we can describe some epistemic advantages of first-person beliefs about seeing.

Note first that John and I are a rough match (modulo the different contents of our beliefs S* and J*) with the immediately preceding example. From before, our closest failed counterparts are equally close. But now we need to consider whether our respective closest counterparts include any big mistakes. Relative to the project of accurately apprehending my subjective situation, the big mistakes are Type A and Type D failure. The closest failed counterparts of John's belief J* do include some big mistakes (third-person Type A and third-person Type D). But the closest failed counterparts (Type B* and Type C*) of my belief S* are not big mistakes. To that extent, my belief is in better epistemic shape.

Coordinately, my belief S* is much safer from severe error. Type A failure (triggered by full-blown matching hallucinations) is a very distant possibility for me. By contrast, third-person Type A failure, as occurring in the "eyes-closed" scenario, is a much closer possibility for John. Turning to Type D, I encounter such failure when I have a matching partial hallucination; this again is a quite exotic possibility. Third-person Type D failure, induced by unfriendly but mundane terrain, is much closer for John.

A "severe-error" safety scorecard is a scorecard that measures subjects' epistemic safety from severe errors. In our context, the severe-error scorecard is a scorecard that is restricted to categories A and D. First-person beliefs about seeing come out as much safer on this scorecard.

We can trace the appearance of the A/D scorecard back to our interest in self-knowledge. This interest yielded assignments of relative severity to our four types of error, and these assignments led to consideration of the A/D scorecard. This scorecard is quite favourable to the first person, and in turn to the naïve-realist externalist project.

Starting again from our interest in self-knowledge, focus on the A/D scorecard can be motivated on slightly different grounds.

Categories A through D represent different ways things can go wrong for John and I as we form beliefs about my subjective situation. Corresponding to each of these ways is a distinctive predication or commitment. Talking about these distinctive predications involves abstraction from beliefs J* and S* as John and I actually hold them. But this abstraction is illuminating. Schematically, the distinctive predications are

(Type A): P sees something.
(Type D): P sees that object.
(Type B*): That object is a dog.
(Type C*): That object is brown.

On their own, the B and C predications do not make claims about my mental life. On their own, they are straightforward claims about one's environment. The B* and C* predications simply make claims about the kind and color of certain objects. Of course, the B* and C* predications do not exist in isolation, they exist as elements of beliefs S* and J*. As parts of these packages, the B* and C* predications do help characterize my mental life.[24] But the B* and C* predications need this packaging --- they need predications A and D --- in order to play this role. By contrast, the A and D predications directly characterize my mental life, without any enabling help from the B* and C* predications. Categories A and D are directly relevant to our examination of self-knowledge. But the relevance of categories B* and C* is entirely due to their connection with A and D.

Another sense in which categories B* and C* are dependent on categories A and D was mentioned earlier. In order to make a B* mistake or a C* mistake, John and I must avoid Type A and D mistakes. We are up for B*/C* evaluation only if we accurately apprehend that I see something (Type A), and only if we accurately pick out (as so) the object that I see (Type D). B*/C* error and success rest on a foundation of A/D success.

Type A and Type D predications form what I will call the "core components" of first-person and third-person beliefs about seeing. These predications have basic mental relevance, while the relevance of predications B* and C* is derivative. Type A and Type D errors are the fundamental types of error that subjects must navigate as they seek to characterize their own visual lives and those of others.

Say that a "core-components" safety scorecard measures the epistemic safety of the core components of first-person and third-person beliefs about seeing. The core-components scorecard is the earlier-identified A/D scorecard. As noted, the first person is well ahead on this scorecard. More explicitly,

> (CORE COMPONENTS) The core components of first-person belief about seeing are much safer than the core components of third-person belief about seeing.

A related reminder is that first-person beliefs are much safer from big mistakes.

We have seen that the relationship between first-person and third-person beliefs about seeing is a textured one, and that our stock of epistemic concepts allows us to make a number of observations about this relationship. As a body, however, these observations are friendly to the naïve-realist cause. Naive realism is compatible with the first person's possession of a sizable, and multi-dimensional, epistemic advantage regarding the nature of his visual experiences.

Naive realism is not the only party which says that factive or success states like seeing are genuine mental states (see also Williamson 2000). One reaction to this outlook parallels a common reaction to externalism about mental content: we accept success states as mental states, but we hold that first-person subjects have no epistemic advantage with respect to their success states (cf. Byrne 2005, pp. 3, 4, n. 3). I oppose this reaction, at least in connection with seeing. I have argued that the core components of first-person belief about seeing are much safer than the core components of third-person belief about seeing. Contrary to natural first impressions, the introduction of success-states does not compromise the epistemic advantage of the first person.

Section V

Almost all philosophers agree that we acquire knowledge of our own minds via methods that are asymmetric to, or "utterly different" from, the methods that we use to acquire knowledge of the minds of others.[25] We might worry that my proposal is not an obvious fit with what I will call "method asymmetry." I have described both first-person and third-person knowledge of seeing in terms of non-inferential perceptual-recognitional abilities. You can perceptually recognize that you see something, and you can perceptually recognize that someone else sees something. My view does entail that at a certain level of description, the methods that first-person and third-person subjects employ in this neighbourhood are parallel. This seems to flout method asymmetry.

However, our apparently troublesome level of description abstracts from important details. Although both the first-person and the third-person employ perceptual-recognitional abilities, these abilities have different structures. I can know that I see a particular dog by looking at the dog, and employing a perceptual-recognitional ability. However, a third-person subject cannot know that I see a particular dog by looking at the dog and employing a perceptual-recognitional ability. Third-party perception of the dog itself --- even if I do see the dog --- is not enough to activate knowledge that I see the dog, even if the third-person subject is conceptually competent and interested in the topic. In order to know, on perceptual grounds, that I see the dog, the third-person subject also needs to see me, and he must do so in a way that makes it possible for him to ascertain my relation to the dog. A third-person subject has the following visual checklist: he needs to see me, and he needs to see the dog. But I only need to see the dog.

Consider the following, very different scenario. I have a lower-deck seat to the baseball game, and you have an upper-deck seat. I have a better view, and this means that I can see details of the unfolding baseball game that you cannot. Suppose that I'm able to see, and know, that a base-runner missed the bag at second base, while you are not in a

position to know this. My lower-deck seat gives me an epistemic advantage. Still, my cognitive relation to the relevant body of fact is not "utterly different" from yours. I simply have a better seat. We can suppose that both seats are behind third base on a line continuous with the baseline from second base to third base. I'm fairly close to the playing field, and you are not. Situated in our respective spots, we both follow the game by watching it. My method for acquiring knowledge of the game is not asymmetric to yours. You are simply at a less advantageous point along the very same dimension.

However, things change when we return to the matter of whether I see a brown dog. Again, in order to know, on perceptual grounds, that I see a brown dog, a third-person subject needs to see me, and he needs to see the dog. But I only need to see the dog. Here there is an important asymmetry in the methods we employ. Here there are clear structural differences between my path to knowledge of the relevant fact and the third-person path. Both parties do employ recognitional abilities. But the lesson to take from this is that the general concept of a recognitional ability subsumes belief-forming methods with clear structural differences. Thus the naïve-realist's appeal to perceptual-recognitional abilities does not compromise the idea that first-person epistemic methods are asymmetric to third-person epistemic methods.

Section VI

I have argued that naive realism, a form of externalism about experience, is compatible with a moderate, substantial conception of first-person privileged access. This conception is defined in terms of the types of first-person epistemic advantage and method asymmetry that I have described in earlier parts of this paper. As noted, my epistemic-advantage thesis is restricted to typical first-person belief and typical third-person belief. In this concluding section, I briefly indicate the sort of case that my "typical" restriction excludes. I will describe one non-perceptual example and then one that involves seeing.

Suppose that Bob dislikes his job, and that he is aware of this attitude. He stays at the job because he has financial responsibilities that he must fulfil. In the workplace Bob makes a concerted and adept effort to conceal his negative attitude towards his position. But suppose an officious co-worker takes a long-term, intense interest in Bob's movements, mannerisms, and workplace conversation. Suppose further that this co-worker is a very astute observer of the human animal. Let's suppose for present purposes that by drawing on long observation and every ounce of his abilities, the co-worker is able to come to know just as well as Bob does that Bob dislikes his job.

Here we have a local collapse of first-person epistemic advantage.[26] But we do not have a counterexample to a first-person epistemic advantage thesis restricted to typical subjects (see Byrne 2005, p. 3, and section II of my present paper). The co-worker's meddlesome nature, his powers of observation, and his lengthy time investment entail that he is not a typical third-person subject. For the project of comparing typical first-person belief and typical third-person belief, the co-worker falls outside the relevant focus group.

Connecting back to our own concerns, it is fairly easy to imagine an analogous character, one who is very good at determining whether or not you see something. Consider a Shoulder-Percher, a being who hovers just above one's shoulder, and who is accordingly quite able to tell when one's eyes are open, and quite able to ascertain one's lines of sight. A Shoulder-Percher's beliefs about what you see are about as safe as your beliefs about what you see. But Shoulder-Perchers are not typical third-person subjects.

When we restrict discussion to typical subjects, first-person epistemic advantage comes to this: if you take a typical first-person subject and a typical third-person subject, and, as a part of this, give them equal resources (e.g., equal levels of rationality and conceptual competence, equal amounts of time), then, due to his first-person status, the first-person subject comes out well ahead, epistemically, on the topic of his own mind.

Note that this conception of epistemic advantage depicts the first-person perspective itself as having a very important epistemic role. The first-person property itself makes a notable epistemic difference.

I have argued that first-person belief comes out well ahead even when the mental state in question is the externally-constituted success state of seeing. The core components of typical first-person belief that one sees something are much safer than the core components of typical third-person belief that something else sees something. The naive-realist theory of veridical experience is compatible with a moderate, substantial conception of privileged access. A more general moral is that success states do not break up the dynamics that many philosophers expect from the phenomenon of knowledge of the mind.*

REFERENCES

Alston, William. 1971: "Varieties of Privileged Access," *American Philosophical Quarterly* 8: 223-241.

-----. 1986: "Internalism and Externalism in Epistemology," reprinted in Alston, *Epistemic Justification: Essays in the Theory of Knowledge*. Ithaca, NY: Cornell University Press, 1989.

Armstrong, David. 1981: "What is Consciousness?," in Armstrong, *The Nature of Mind*. Brighton, Sussex: Harvester Press, 1981.

Bergmann, Michael. 1997: "Internalism, Externalism, and the No-Defeater Condition," *Synthese* 110: 399-417.

Boghossian, Paul. 1989: "Content and Self-Knowledge," *Philosophical Topics* 17: 5-26.

-----. 1997: "What the Externalist Can Know A Priori," *Proceedings of the Aristotelian Society* 97: 161-175.

Brown, Jessica. 1995: "The Incompatibility of Anti-Individualism and Privileged Access," *Analysis* 55: 149-156.

-----. 2000: "Reliabilism, Knowledge, and Mental Content," 100(1): 115-135.

Burge, Tyler. 1988: "Individualism and Self-Knowledge," *Journal of Philosophy* 85: 649 -633.

Byrne, Alex. 2005: "Introspection," *Philosophical Topics* 33(1). Available online at http://mit.edu/abyrne/www/introspection.pdf.

Comesana, Juan. 2005: "Unsafe Knowledge," *Synthese* 146: 395-404.

Conee, Earl and Richard Feldman. 2001: "Internalism Defended," in *Epistemology: Internalism and Externalism*, ed. Hilary Kornblith. Oxford: Blackwell.

DeRose, Keith. 2004: "Sosa, Safety, Sensitivity, and Skeptical Hypotheses," in *Sosa and His Critics*, ed. John Greco. Malden, MA: Blackwell.

Drestke, Fred. 1995: *Naturalizing the Mind*. Cambridge, MA: MIT Press.

-----. 1999: "The Mind's Awareness of Itself," *Philosophical Studies* 95: 1-22. Reprinted in Dretske, *Perception, Knowledge and Belief*. Cambridge: Cambridge University Press, 1999. Page references are to the reprint.

-----. 2003a: "How Do You Know that You're not a Zombie," in Gertler (ed.) 2003.

-----. 2003b: "Externalism and Self-Knowledge," in Nuccetelli (ed.) 2003.

Gertler, Brie. 2000: "The Mechanics of Self-Knowledge," *Philosophical Topics* 28(2): 125-146.

----- (ed.). 2003: *Privileged Access: Philosophical Accounts of Self-Knowledge*. Aldershot: Ashgate Publishing.

Goldman, Alvin. 1999: "Internalism Exposed," *Journal of Philosophy* 96: 271-293.

Haddock, Adrian, and Fiona Macpherson (eds). 2008: *Disjunctivism: Perception, Action, and Knowledge*. Oxford: Oxford University Press.

Hawthorne, John. 2007: "A Priority and Externalism," in *Internalism and Externalism in Semantics and Epistemology,* ed. Sanford Goldberg. Oxford: Oxford University Press.

Hawthorne, John, and Karson Kovakovich. 2006: "Disjunctivism," *Proceedings of the Aristotelian Society Supp. Volume* 80: 145-183.

Heil, John. 1988: "Privileged Access," *Mind* 97: 238-251.

Loar, Brian. 1990/1997: "Phenomenal States," in *Philosophical Perspectives* 4: 81-108. Revised version in *The Nature of Consciousness*, eds. Ned Block, Owen Flanagan, and Güven Güzeldere. Cambridge, MA: MIT Press.

Lycan, William. 1996: *Consciousness and Experience*. Cambridge, MA: MIT Press.

-----. 1999: "Dretske on the Mind's Awareness of Itself," *Philosophical Studies* 95: 125-133.

Ludlow, Peter and Norah Martin (eds.). 1998: *Externalism and Self-Knowledge*. Stanford, California: CSLI Publications.

Manley, David. 2007: "Safety, Content, Apriority, Self-Knowledge," *Journal of Philosophy* 104: 403-423.

Martin, Michael. 2004: "The Limits of Self-Awareness," *Philosophical Studies* 120: 37-89.
-----. 2006: "On Being Alienated," in *Perceptual Experience*, eds. Tamar Szabo Gendler and John Hawthorne. Oxford: Oxford University Press.
McDowell, John. 1994: *Mind and World*. Cambridge, MA: Harvard University Press.
McKinsey, Michael. 1991: "Anti-Individualism and Privileged Access," *Analysis* 51: 9-16.
Millar, Alan. 2007a: "What the Disjunctivist is Right About," *Philosophy and Phenomenological Research* 74: 176-198.
-----. 2007b: "The State of Knowing," *Philosophical Issues* 17: 179-196.
------. 2008: "Perceptual-Recognitional Abilities and Perceptual Knowledge," in *Disjunctivism: Perception, Action, and Knowledge*, eds. Adrian Haddock and Fiona Macpherson. Oxford: Oxford University Press.
Nozick, Robert. 1981: *Philosophical Explanations*. Oxford: Oxford University Press.
Nuccetelli, Susan (ed). 2003: *New Essays on Semantic Externalism and Self-Knowledge*. Cambridge, MA: MIT Press.
Ryle, Gilbert. (1949): *The Concept of Mind*. London: Hutchinson.
Sainsbury, Mark. 1997: "Easy Possibilities," *Philosophical and Phenomenological Research* 57: 907-919.
Sawyer, Sarah. 1999: "An Externalist Account of Introspective Knowledge," *Pacific Philosophical Quarterly* 80: 358-378.
Siegel, Susanna. 2004: "Indiscriminability and the Phenomenal," *Philosophical Studies* 120: 90-112.
Shoemaker, Sydney. 1988: "On Knowing One's Own Mind." *Philosophical Perspectives* 2: 183-209. Page references to reprinting in Shoemaker 1996.
-----. 1994: "Self-Knowledge and "Inner Sense," *Philosophy and Phenomenological Research* 54: 249-314. Page references to reprinting in Shoemaker 1996.
-----. 1996: *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.
Sosa, Ernest. 1999: "How to Defeat Opposition to Moore," *Philosophical Perspectives* 13: 141-154.
-----. 2004: "Relevant Alternatives, Contextualism Included," *Philosophical Studies* 119: 35-65.
Tye, Michael. 2000: *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.
Williamson, Timothy. 2000: *Knowledge and Its Limits*. Oxford: Oxford University Press.

---

[1] Recent discussions of naive realism, which tend to focus on "disjunctivist" developments of this theory, include Martin 2004, 2006, Siegel 2004, Hawthorne and Kovakovich 2006, and the papers collected in Haddock and Macpherson (eds.) 2008. The topics of introspection and self-knowledge are not absent from these and other relevant papers. The familiar idea that there are hallucinations that we cannot distinguish from veridical experiences has an influential presence in the recent work. I suggest that positive naive-realist proposals about first-person knowledge of one's experience have been comparatively underexplored.

[2] There is a large and well known debate concerning the relationship between externalism about mental content and privileged access. Important papers in this discussion include Burge 1988, Heil 1988, McKinsey 1991, Brown 1995, and Boghossian 1997. Recent collections on the topic include Ludlow and Martin (ed.) 1998 and Nuccetelli (ed.) 2003. The recent debate has focused primarily on first-person knowledge of the contents of thought and belief. In the present paper I explore connections between a form of externalism about visual experience and some ideas associated with privileged access.

[3] One privileged-access idea that I do not consider in this paper is that first-person knowledge of one's mind is a priori. My theory of first-person knowledge of experience does give this knowledge an empirical basis, but here I do not pursue the resulting tension between my position and an a priori conception of self-knowledge. I believe that empirical accounts of first-person knowledge of experience can be motivated on grounds independent of naive realism. I pursue this issue in a companion paper (in preparation).

[4] For some discussion of this distinction, see Goldman 1999 and Conee and Feldman 2001. Earlier work on the distinction includes Alston 1986 and Bergman 1997.
[5] For discussion of recognitional abilities, see, e.g., McDowell 1994 (ch. 3), Millar 2007a, 2007b, 2008. Some writers in the recent discussion of phenomenal concepts also appeal to recognitional abilities (Loar 1990/1997, Tye 2000). Also, Sarah Sawyer (1999) makes use of recognitional abilities, and another resource in my present paper, epistemological modal considerations, in defense of externalism in the privilege-access debate. But Sawyer's concerns are quite different from mine. First, Sawyer's topic is first-person knowledge of thought content (1999, p. 359), rather than first-person knowledge of one's perceptual experience. Second, Sawyer's aim is to argue that externalism about mental content is compatible with the apparent a priori character of our knowledge of the contents of our thoughts (pp. 370-371). For this reason, the type of recognitional ability that Sawyer describes is explicitly not a *perceptual* recognitional ability. By contrast, I embrace an empirical account of one type of self-knowledge, and I try to maintain compatibility with two other privileged-access ideas, first-person epistemic advantage, and method asymmetry.
[6] See Shoemaker 1988, 1994.
[7] For earlier use of this helpful phrase, see Gertler 2000.
[8] Dretske 1995; see also Tye 2000. Dretske's 1999 theory of experiential self-knowledge also assigns a role to outer perception (1999, p. 168), but in this paper Dretske does not refer to connecting beliefs or employ the notion of displaced perception. Returning, for our purposes, to this notion, for Dretske displaced perception provides only a limited form of self-knowledge. It allows us to know the contents of our experiences --- to know *what* we experience --- but it does not allow us to know *that* we have experiences (1995, p. 57-58). Dretske elaborates this somewhat skeptical stance in his 2003a. For a note of optimism, however, see Dretske's 2003b, p. 142 n. 6.
[9] Dretske 1995, p. 42.
[10] Aydede 2003, pp. 57-58. For a very brief statement of the same doubts about Dretske's theory, see Lycan 1999, p. 132 n. 4.
[11] For more discussion of the environmental/modal dimension of recognitional abilities, see Millar 2008. In the present paper, modal considerations play a large role in sections II-IV. In his work Millar tends to talk in terms of environments (2008, p. 335), whereas I will speak of possible worlds, but I don't think this reflects a large difference in outlook. Another minor difference with Millar is that in his work he elucidates the modal profiles of recognitional *abilities*, but I will examine the modal profiles of token *beliefs*.
[12] All page references to Byrne 2005 are to the online version available at http://mit.edu/abyrne/www/introspection.pdf.
[13] I will often write of typical subjects rather than typical situations. This is consistent with the quotation from Byrne in the main text of this paper. I assume that if a situation contains at least one relevantly atypical subject, then it is not a typical situation in the sense that Byrne intends. The possibility of formulating epistemic-advantage theses with "normal" restrictors is noted in Alston 1971. The current restriction to typical subjects is in the same spirit.
[14] For a recent, more ambitious statement of first-person epistemic advantage, see Gertler 2000. The claim in Gertler's paper is that all rational subjects have, with respect to the contents of their own occurent thoughts, an epistemic advantage over all possible third-person subjects (pp. 126, 131). We should note two points about Gertler's paper. First, Gertler's epistemic-advantage thesis is explicitly restricted to first-person knowledge of the contents of one's thoughts, not their ingredient attitudes (p. 126). Byrne's more moderate thesis addresses first-person knowledge of attitude *and* content (Byrne 2005, p. 3). Second, there is some evidence that Gertler regards her strong thesis as less plausible as applied to first-person knowledge of attitudes (Gertler 2000, pp. 126, 145 n. 17). If we assume for our present purposes that Gertler's claim about thought content is tenable, then we should expect a considerably weakened privileged-access thesis to apply to a more robust type of self-knowledge, knowledge of attitude and content. And that, in Byrne, is exactly what we find.
[15] For discussion of safety, see, e.g., Sainsbury 1997, Sosa 1999, Williamson 2000, Hawthorne 2007, Manley 2007.
[16] For an earlier discussion of first-person epistemic advantage that makes similar use of the concept of safety, see Byrne 2005.
[17] To the extent that we need a closeness threshold in this paper, I will let it be set by our judgments about the closeness of particular cases, rather than try to provide an independent specification.
[18] See DeRose 2004, pp. 33-34; Sosa 2004, pp. 44-45 (and n. 22); Manley 2007, p. 407.

[19] For an early statement of this point, see Nozick 1981, p. 179. For subsequent discussion, see, e.g., Williamson 2000, pp. 128, 149; Sosa 2004; Comesana 2005, pp. 396-397; Hawthorne 2007, Manley 2007. My main-text basketball example is similar to an example in Hawthorne 2007.
[20] See also Sainsbury 1997, Brown 2000, and Hawthorne 2007.
[21] For brief discussion of typicality, see section VI.
[22] Note that the applicability of "ersatz dogs" talk does not entail that my counterpart thought refers to anything. My counterpart thought involves a failed demonstrative.
[23] For brief endorsement of the idea that both modal proximity and severity of failure are epistemically significant, see Williamson's discussion of reliability and "degrees of danger" (2000, p. 124).
[24] The "characterization" I have in mind here is not a success notion.
[25] The quotation is from Shoemaker 1988, p. 25. The main philosophical opponent of method asymmetry is Gilbert Ryle (1949).
[26] This is not a collapse of what I call method asymmetry (for discussion of method asymmetry, see section V). I'm not suggesting that Bob and his co-worker use the same methods to determine that Bob does not like his job. I'm just suggesting that the two subjects, in this case, could be on an epistemic par.

[*] I presented earlier versions of this material at two meetings of the Metaphysics Group at Nottingham. For helpful feedback, I would like to thank colleagues who participated on those occasions, including Stephen Barker, Mark Jago, Harold Noonan, Philip Percival, and Jonathan Tallant. Special thanks to Carrie Jenkins, whose comments at the Group led to an overhaul of section IV; and to Daniel Nolan, for his instrumental comments on a late draft of the paper, and for several helpful conversations.