Knowing Your Own Beliefs

Eric Schwitzgebel
Department of Philosophy
University of California at Riverside
Riverside, CA 92521-0201

eschwitz at domain: ucr.edu

June 16, 2011

Knowing Your Own Beliefs

How do you know your own beliefs? And how well do you know them? The two questions are related. I'll recommend a pluralist answer to the first question. The answer to the second question, I'll suggest, varies depending on features of the case.

1. Four Approaches to Self-Knowledge of Belief.

Self-scanning. Shaun Nichols and Stephen Stich (2003) say this: You have in your mind a functionally defined "belief box". To believe some proposition P is just to have a representation with the content "P" in the belief box. You also have a monitoring mechanism that can scan the contents of the belief box. Normally, you come to know what you believe by deploying that scanner: creating a new belief in the belief box, a belief with the content "I believe that P". Self-scanning accounts admit of many possible variations and complications (e.g., Armstrong 1968, 1981, 1999; Lycan 1996; Goldman 2006), but the basic idea is that people have one or more interior monitors or scanners that detect the presence of beliefs and produce, as output, beliefs about those beliefs (or judgments about those beliefs, or representations of those beliefs).

Transparency. Here's a very different approach: When asked whether I believe that P, I don't scan myself to discover the presence or absence of some pre-existing interior state of belief that P; rather, I think about, I consider, P-relevant facts about the outside world. I consider whether P is true. In Gareth Evans's (1982) famous example, if I am asked whether I believe there will be a third world war, I think about *whether there will be a third world war.* I turn my gaze, as it were, outward not inward. This is sometimes called a "transparency" approach to

self-knowledge, and it has been developed in very different ways by, for example, Fred Dretske (1995), Michael Tye (2000), Richard Moran (2001), Dorit Bar-On (2004), Alex Byrne (2005), Nishi Shah and David Velleman (2005), and Robert Gordon (2007). An *expressivist* approach to transparency might involve the self-attribution "I think P" to arise spontaneously from reflection on the P-ishness of the world much as an utterance like "ow!" might arise spontaneously from dropping an object on one's toe. An *inferentialist* approach to transparency might involve the self-ascriptive conclusion "I believe that P" to follow straightaway from "P" much as the conclusion "P or Q" follows straightaway from "P" – with no more self-scanning required in the former case than in the latter.

Let me draw attention here to an infrequently remarked-upon division within the transparency literature between *topic-shifting* and *self-judgment* approaches. Transparent reflection is topic-shifting if it involves no judgment about one's own mental states – if, for example, one hears the question "Do you think that P?" merely as a variant on the question "Is P the case?", so that the answer "yes" or "I think so", though in the context literally self-ascriptive, no more involves a judgment about one's own mind than does any ordinary assertion about the world. Self-judgment approaches, in contrast, treat the transparency process as creating, as its end-product, a judgment (representation, belief, whatever) about one's own mind.

Theory theory. According to theory theorists, like Daryl Bem (1967), Alison Gopnik (1993a-b), Krista Lawlor (2008), and Peter Carruthers (2009, forthcoming), although self-knowledge might *seem* to be a simple, straightforward matter, actually it is the consequence of a process of theorizing about – or (if "theorizing" sounds too heavy-handed) interpreting or inferring – mental states to which one has no direct access. Just as to the chess master, the weakness of a chess position might seem to be simply visually given, despite complex layers of

theory or inference underneath, so also to the skilled self-theorist, the presence of a belief might seem to be simply introspectively given, despite (potentially) complex layers of theory or inference underneath. How, on such a view, do I know that I believe it will rain? Well, I see myself carrying an umbrella, I notice that I have said (in inner or outer speech) "it looks like rain", I detect a mental image of myself getting wet, I see dark clouds and I know that reasonable people tend to think it will rain when they see dark clouds. I'm a practiced enough folk psychologist to swiftly infer from all this evidence that I believe it will rain – just I would infer the same about someone else if I had the same set of evidence.

Partial constitution. John Heil (1988), Sydney Shoemaker (1996, 2009), and others have suggested that it's just *part of what it is* to believe that P that you are disposed to self-attribute belief that P. (I assume throughout this essay, as partial-constitution folks also assume, that the believer has a mature set of human background capacities and concepts, including the concept of "belief". Some maneuvering is necessary to reconcile partial constitution views of belief ascription to infants and non-human animals.) Plausibly, on a dispositionalist or functionalist or interpretationist approach to belief, part of what it is to believe that the king is fat is to be disposed, in the right circumstances, to say things like, "I think the king is fat". The belief that P and the disposition to self-attribute the belief that P are thus not ontologically distinct in the way that seems to be presupposed by the other approaches. There is not one state, the belief, and then a separately existing mechanism or process by means of which we self-attribute that belief.

Each of these four approaches – self-scanning, transparency, theory theory, and partial constitution – admits of variants, and the views can be brought together in muliple ways. The view I present in this essay is a partial-constitution view, but one that recognizes substantial truth in the other three approaches. This view is grounded in two background positions: a liberal

dispositionalism about the nature of belief and a commitment to the idea that self-judgments – like most of our judgments – tend to flow from a complex multiplicity of competing and co-operating processes.

2. The Metaphysics of Belief and Partial Constitution.

On my view, to believe some proposition P – to believe, let's say, that it will probably rain on Thursday – is to match, to a sufficient extent, a particular *dispositional profile* – to possess, that is, a cluster of dispositions that people do, or would, associate with the belief that P. Although dispositional approaches, historically, sometimes emphasized only behavioral dispositions (e.g., Braithwaite 1932-1933; Marcus 1990), I recommend what we might call *liberal dispositionalism.* On liberal dispositionalism, the range of dispositions relevant to belief ascription includes not only behavioral dispositions but also dispositions to enter various phenomenal and cognitive states. Thus, to believe that it will probably rain on Thursday is to be disposed not only to say, in appropriate contexts, "It will probably rain on Thursday" and to propose days other than Thursday for the company picnic, but also to be prone to say to oneself in silent inner speech that it will probably rain Thursday, to feel surprise if one wakes Thursday morning to find it clear and sunny, to be apt to infer that it will probably rain sometime this week should that question arise, and to start to think of Thursday as an indoors type of day. Any person who has this general dispositional profile believes; any person who lacks this dispositional profile fails to believe. Internal representations, brain structure, and the like are relevant only to the extent that they underwrite possession of this dispositional profile; and "mad belief" – i.e., believing but not possessing any of the behavioral, phenomenal, or cognitive dispositions characteristic of belief – is conceptually impossible (Ryle 1949; Price 1969;

Schwitzgebel 2002, 2006/2010, forthcoming; Hunter forthcoming; contrast Lewis 1980 on "mad pain").

One major complication for dispositionalism about belief is that the relevant dispositions hold only *ceteris paribus* – all else being equal, or all else being right. Our Thursday-rain believer will not be disposed to avoid planning the company picnic on Thursday if she thinks the company would enjoy a picnic in the rain or if she wants to sabotage the picnic. And although she might be disposed, generally speaking, to draw obvious consequences from P should the occasion arise, we all fail to draw obvious consequences from time to time. Such facts don't seem to make it any less true that she believes. One might attempt to address this issue by turning the antecedents of the disposition-specifying conditionals into complicated conjunctions: If she wants the picnic to succeed and if she thinks that the company would not like a picnic in the rain and if she knows the picnic is to be outdoors, etc., then she won't plan the picnic for Thursday. The challenge, then, is to non-circularly complete the "etc.". Another strategy, however – I think a better strategy – is simply to accept the irreducibly *ceteris paribus* nature of the dispositions constitutive of belief. Most scientific generalizations, too, only hold *ceteris paribus* – a point emphasized by Nancy Cartwright (1983). This doesn't make them useless or bad science. Objects in free fall accelerate toward the earth at about 10 meters per second squared – unless their shape and mass create significant air resistance, unless they are in a strong magnetic field and made of iron, unless the earth has lost half its mass due to a horrible collision with another celestial body, etc. We skillfully disregard such exceptions when they are irrelevant to the case at hand, and when they are relevant, we skillfully conditionalize: Objects in free fall accelerate at about 10 meters per second squared, assuming air resistance is negligible. So also, I

suggest, with the dispositions constitutive of believing. (Indeed, dispositions are plausibly seen as modal generalizations.)

The core idea of liberal dispositionalism about belief is that to believe that P is nothing more or less than to be disposed to act and react – not just outwardly but also in one's phenomenology and patterns of cognition – shall we say, belief-that-P-ishly. What is it to act and react belief-that-P-ishly? It is, I suppose, to act and react as though P were the case. But what is *that?* What unifies all the various things that appear to be involved in acting and reacting as though P were the case – the picnic planning, the disposition to be surprised by a sunny sky in two days, a certain stream of inner speech? I recommend that rather than trying to find a substantive unifying thread (such as acting and reacting in ways that would advance one's goals if P were true) we go sociological: To believe that P is to act and react, and be otherwise disposed, in ways that ordinary people would regard as characteristic of believing that P. This might sound circular, but it's not. It is to ground metaphysics in folk psychology: There are some patterns in the world that ordinary belief ascribers have glommed on to, patterns that drive their aptitude to regard certain dispositional tendencies as characteristic of believers, including assertion, inference, implicit assumption, lack of surprise, and dependence on its truth in making plans. Assuming there is something to that folk-recognized pattern, we can build a metaphysics parasitic on it – at least until we arrive at something better, if ever we do. (We can also "rigidify", allowing us to apply our sense of the patterns even to hypothetical cases in which folk psychology is different than it in fact is.)

For the purposes of this essay, though, it doesn't really matter if you accept my particular species of liberal dispositionalism about belief. The partial constitution view will follow from any dispositional approach to belief as long as *among the dispositions constitutive of believing*

that P is the disposition to self-ascribe belief that P. Part of acting and reacting belief-that-P-ishly, presumably, on almost any broad-based dispositional account, is to say things like "I believe that P". If so, the disposition to self-ascribe belief that P is not entirely ontologically distinct from the belief that P. *Part of what it is* to believe that P is to be ready to attribute belief that P to oneself when the occasion arises. (Parallel remarks apply to functionalist accounts of belief.)

This might sound pretty good for self-knowledge. Advocates of partial constitution views, like Shoemaker, tend to emphasize the positive epistemic consequences. But here's the catch: On any plausible account, the self-ascriptive disposition will be only a *small part* of the overall dispositional profile constitutive of belief that P. Someone who lacks that part of the profile might still match well enough the remainder of the profile that it's still reasonable to describe her as believing that P. Consider an analogy. Part of being courageous is being disposed not to flee from battle. Being disposed not to flee from battle is partly constitutive of the character trait of courage. And yet someone who lacks that disposition – someone who is disposed to flee from battle – might still accurately, or accurately enough, be described as courageous, if she has enough of the other dispositions constitutive of courage. If she will stand up for justice against authority, risk her life to save a person in need, keep her cool in an emergency, we might still rightly enough say she is courageous, despite what we know about her dispositions in battle – especially if her dispositions in battle are not really relevant in the context of ascription. For complex, "multi-track" dispositional properties, we can sometimes disregard *outlier dispositions*. Accordingly, if someone is behaviorally, phenomenally, and cognitively disposed perfectly belief-that-P-ishly across the board, lacking *only* the tendency to self-ascribe

that belief, we might well enough in most circumstances go ahead and say that she believes that P.

How empirically likely it is for a person to have a single outlier disposition of that sort is another question. More commonly, I suspect, as a matter of empirical fact, when someone deviates from the dispositional profile constitutive of belief that P, he will tend to do so in multiple ways. But that doesn't mean, necessarily, that it will be completely appropriate to deny him the belief that P, or that he will determinately fail to believe that P. There will, I suggest, be a range of *in-between* cases – cases in which the subject has enough of the dispositional profile that it would be misleading simply to say that he does not believe that P, but in which he also deviates enough from the profile that it would be misleading simply to say that he does believe that P. (I have discussed cases of this sort extensively in several essays: Schwitzgebel 1999, 2001, 2002, 2010.) Analogizing again to personality traits: If a person is courageous in some ways and not courageous, or even cowardly, in others, it might not be quite right either to simply ascribe or to simply deny the trait of courage. It might be a mixed-up, in-between case in which the best answer to a question about whether he is courageous is to refuse to label him either way and instead to specify with detail: Well, when it comes to risking his life, he is not courageous, but he is courageous is standing up to authority and in risking livelihood, welfare, and the condemnation of others to fight for what he thinks is right. The move of refusing to answer with a simple attribution or denial and instead specifying detail is common also in-between cases of simpler vague predicates: When asked "is he skinny?" of an in-between case, instead of saying yes or no the more ordinary thing to say would be something like: "Well, he's thinner than Vijay but not as skinny as Steve" or "Well, he's 5'11" and 175 pounds". In-between cases of belief, where subjects possess a substantial portion of the dispositional profile but not enough for belief

ascription to be straightforwardly accurate across conversational contexts include, for example: cases of gradual learning and forgetting, where the subject has a very tentative and inconsistent grip on something; cases in which the subject consistently implicitly assumes that P is the case but will sincerely deny the truth of P when asked explicitly, such as in implicit racism; cases of self-deception; and cases of conceptual or referential confusion, such as Kripke's (1979) famous case of Pierre, who doesn't realize that "Londres" and "London" refer to the same city. And in such in-between cases we normally don't find good self-knowledge: People are usually not well attuned to their patterns of dispositional splintering. The implicit racist normally has no specially privileged knowledge of her implicit racism; the confused or self-deceived person rarely knows the extent of her confusion or self-deception.

Liberal dispositionalism about belief make it plausible, at least in the abstract, that there should be *some* in-between cases of belief: If to believe is to have a certain cluster of dispositions or to be prone to act and react belief-that-P-ishly, then it seems that we could match that cluster or have that proneness either not at all, somewhat, to a substantial degree, or entirely. Somewhere along this spectrum will be vague or in-between cases in which neither simple attribution nor simple denial of the belief is quite right. Similar in-betweenness follows too, I think, from other approaches to the metaphysics of belief, such as multi-track functionalist or interpretationist approaches. A complex functional role may be *partly* filled; an interpretation may be *to a certain extent* apt. If we build our theories on clean cases alone, we risk forgetting this.

In sum: To believe, according to the view I will assume for the remainder of the essay, is to possess a cluster of behavioral, phenomenal, and cognitive dispositions. Among those dispositions are self-ascriptive ones. Therefore, self-ascription of the belief that P is partly

constitutive of believing that P. Nonetheless, it remains an empirical question how reliably the dispositions constitutive of the belief that P will travel together. Self-ascriptive dispositions may splinter away from the rest, or from many of the rest, creating mismatches between what one believes and what one self-ascribes. In Section 4, I will present some cases.

3. Mechanisms of Self-Knowledge.

One thing to notice about the dispositional account just offered is that it is silent about mechanisms. Although someone who fully matches the dispositional profile for believing that P will necessarily be disposed to self-ascribe belief that P, the dispositional account says nothing about in virtue of what she will possess her self-ascriptive dispositions, says nothing about what cognitive processes underwrite them. I suggest that there are several, perhaps many (depending on how one counts) cognitive processes that can underwrite such dispositions, including all the types of process described in Section 1, in various forms and guises, often acting co-operatively or competitively.

I ask Laura if she believes that there is anything wrong with gay men having consensual sexual intercourse, and she answers that she sees nothing wrong with it. What processes might drive Laura's self-ascription? Why not: lots? She might answer the question in part by thinking about whether there really is anything wrong with gay men having consensual sexual intercourse – and in doing so she might partly be considering the matter afresh (especially if she hasn't heard the question put to her in exactly those words before) and she might partly be calling up the moral facts (or putative moral facts) from memory, much as a schoolchild might call up California's capital from memory. (However, even when we seem simply to be calling from memory we are often half re-evaluating – attuned at least implicitly to recent relevant changes

and conditions.) Laura's self-ascription of that attitude might also be partly driven by her general conception of herself as liberal and her memory of having explicitly endorsed similar propositions in the past. She is motivated to be self-consistent and non-homophobic – and that motivation, as I am imagining it, is not just evidential: Although Laura might regard the fact that she has asserted P in the past as evidence both of P's truth and of her belief that P, she is also, beyond that, motivated to be self-consistent over time; although she might think that non-homophobes generally have better command of the facts, her desire not to be homophobic colors her self-ascriptional dispositions for more than just that reason. Laura's self-ascription, though sincere, might nonetheless reflect something of her expectations about the audience: She might be disposed to judge herself a simple liberal on the issue if asked by me, while at the same time she is disposed to self-attribute a more nuanced view, equally sincerely, were she to be asked the same question by her conservative Christian uncle. She might visually imagine gay sex and have an emotional or aesthetic reaction that influences her self-ascription, one way or the other (and not all influences have to point in the same direction). There might be a simple association between thinking about gay rights and an impulse to make affirmative utterances, either self-ascriptively or non-self-ascriptively. Laura's affirmation, whether explicitly self-ascriptive or not, might help to reinforce the opinion she expresses. Since our minds are massively interconnected, there might also be some relatively direct causal connection, or set of connections, invisible to introspection, between the brain states or brain processes or cognitive states or cognitive processes involved in self-ascription and the brain states or brain processes or cognitive states or cognitive processes involved in having the other dispositions constitutive of believing that there is nothing wrong in consensual sex among gay men.

Laura *could* perhaps simply think about P and then self-ascribe belief on that basis alone, as suggested by the self-judgment version of the transparency view. But to the extent that she is reaching a judgment about her own mind and not simply shifting the topic only to P itself, her self-conceptions will, it seems, both properly and almost inevitably come into play. And unless she is reaching a judgment at least in part about her own mind, self-knowledge is not really at issue. A two-and-a-half year old, not even possessed of the concept of belief (at least in mainstream developmental psychological opinion) can answer the question "Do you think that P?" by interpreting it as something like "blah-blah-blah P?" and thus avail himself of the topic-shifting transparency strategy described in Section 1; that's not self-knowledge. Self-conceptions *properly* come into play in self-ascription because – as Victoria McGeer has emphasized (2007a&b; McGeer and Pettit 2002) – part of what it is to be a believer, at least for normal, mature human beings, is to be someone who strives to be consistent and interpretable over time, acting, opining, and self-regulating in accord with folk psychological norms. Relatedly: If I'm tilting toward P and "I believe that P" seems to follow, I don't have to do modus ponens and self-attribute P; if the latter sits uncomfortably with my self-conception or past avowals I can, and often should, do modus tollens instead and rethink P itself. This is transparency in reverse.

And even transparency theories can plausibly operate through different cognitive mechanisms. For example, I might *infer* "I believe that P" from P (e.g., Dretske 1995; Byrne 2005 – though neither author appears to be entirely comfortable with calling this "inference" in the ordinary sense). Alternatively, thinking about P might promote in me a self-expressive impulse, which I disinhibit, to utter, in inner or outer speech, "Yeah, I think that P!" (e.g., Bar-

On 2004; Gordon 2007). Do we need to say that one or the other of these procedures is the single true story?

Laura *could* perhaps simply note that she is saying "P" to herself in inner speech, could recall her previous avowals of P, could note patterns in her behavior that seem to suggest a P-ish attitude, and reach the conclusion that she believes that P on that basis alone. But that would seem to be an unusual, limiting case. If transparency strategies are cognitively possible; and if they are, as they seem to be, fast and easy; and if they have at least some tendency to generate correct answers, it would seem odd not to frequently avail ourselves of them. (We might also consider: "I want X" from X is good, "I'm afraid of X" from X is dangerous, "I hate X" from X is horrible, etc. – strategies of varying reliability.) And transparency doesn't seem to be the type of process normally invoked by theory-theorists. Theory theorists rightly emphasize the often underappreciated fact that our self-ascriptions are importantly *influenced* by self-observation of speech and other behavior. But it by no means follows that theoretical inference or self-interpretation the *only* basis of belief self-ascription. Pluralism could be true.

There might also be relatively direct self-scanning mechanisms in the mind: The brain is massively interconnected, and different cognitive subsystems presumably track each other in a variety of ways. One way this might work is as follows: Just as a memory trace can directly influence (directly enough, at least, for current purposes) my history-related judgments and speech, so also, in much the same way, the very same memory trace, or perhaps a self-ascriptive version of it (or why not both?), could presumably directly influence my self-ascriptive judgments and speech. I'm not sure that we need to think of such a process in terms of scanning and retrieval, exactly, but it the direct effect of memory traces is something neither transparency theory nor theory theory appears to capture and which seems to be more or less the sort of thing

that self-scanning theorists have in mind. Contra theory theory, a self-scanning process needn't avail itself of any evidential ground, in behavior or phenomenology, for the self-ascriptive judgment. Contra transparency views, a self-scanning process needn't be derivative of the process by which I produce the judgment that P: It might be prior, or simultaneous but independent, or simultaneous and symmetrically co-dependent. Explicitly self-ascriptive memory traces perhaps reveal the possibility most clearly: Just as I can store and retrieve (as it were) "Plato taught Aristotle", it seems I can store and retrieve "I believe that Plato taught Aristotle". It would be a strange incapacity if the latter were not the sort of thing that I could directly remember but something, rather, that always had to be reconstructed from reflection on the fact that Plato taught Aristotle or from the evidence of my behavior and imagery. This isn't to deny pluralism, of course. I see no reason to think that straight memory retrieval, or any other scanning-type processes, would tend to operate solo in self-ascription. In fact, ordinary non-self-ascriptive recall rarely operates simply by dumb retrieval, but is always or almost always half-reconstructive in light of existing background knowledge, schemas, and environmental information (which, however, makes the storage-and-retrieval metaphor somewhat problematic: Bartlett 1932; Roediger 1980; Sutton 1998; see also Hurlburt and Schwitzgebel 2011).

Shall we run another example? I chose the gay sex example to make vivid the variety of processes driving self-ascription. But the same pluralism is plausible for more pedestrian examples too, such as the belief that it will probably rain on Thursday. At least to the extent our subject – LeToya, let's call her – is reaching a self-judgment (rather than doing a transparent topic shift), her self-conception will be relevant to her assessment of what she believes: She thinks of herself as pessimistic; in fact she is somewhat proud of being a pessimist. That the Thursday-rain belief is consonant with her general pessimism (since the company picnic *was*

planned for Thursday) adds a layer of comfort to that self-ascription; LeToya would think twice about self-ascribing an optimistic-seeming belief. She remembers that she warned the secretary yesterday (Monday) about the forecasted storm. That reinforces her self-ascription: She has taken her prescient, pessimistic stand! And the evidence continues to support her view: An hour ago, she saw the looming low pressure zone on the internet. She re-envisions the weather map. She pictures wet potato salad and a wet plant manager. "You guys are gonna get soaked" just sounds right on her lips. LeToya can't resist saying it, and once she says it, she knows she is even more on the line for it. At the same time, perhaps, her pessimism is – as she half-realizes – partly a pose: If someone were to confidently express more pessimism about the matter than she, she would suddenly find herself expressing restrained optimism. She has seen that pattern in herself before. She'll take no bets.

Why, I wonder, would anyone be attracted to a monolithic theory? The mind is so delightfully multi-faceted, multi-capacitied, and complex!

Let me conclude this section with one clarification: I have spoken throughout of self-ascription or self-ascriptive judgment. However, it perhaps bears mention that just as the belief that P involves a complex cluster of dispositions, so also does the higher-order belief that I believe that P. Other dispositions partly constitutive of the belief that I believe that P include: the disposition to have a "feeling of knowing" about P even when P doesn't immediately pop to mind (e.g., I might feel I know my college roommate's last name even as I am failing to come out with it); the disposition to feel a certain sort of surprise if I were to discover not-P, surprise not only at the unexpectedness of not-P but a further dimension of surprise as well, reflecting the fact that I thought I knew that P; a general sense of not being ignorant about P-related topics, which can drive my choices of what to write or talk about; and the disposition *not* to feel

surprised or like I have learned something about myself when I hear myself endorsing P (a disposition we sometimes lack, as reflected in the bon mot about not knowing what I think until hearing what I have to say). The dispositions constitutive of higher-order belief can splinter, just as the dispositions constitutive of simple, lower-order belief can splinter.

4. Good Self-Knowledge, Poor Self-Knowledge.

It is, I suggested, an empirical question how reliably the disposition to self-ascribe belief that P tends to align with the remaining dispositions constitutive of belief that P. Here's my guess: The relationship will be good to the extent that the belief (or other attitude) meets three conditions: (1.) Possession of the belief is normatively neutral, in the sense that having the belief reflects, in the subject's own view, neither poorly nor well upon her, except insofar as believing truths and not believing falsehoods reflects well on her. (2.) The belief is straightforwardly connected to observable behavior. And (3.) the dispositions most relevant to the ascription of the belief that P are inward and outward judgments or statements that P is the case. Conversely, the relationship between self-ascriptive dispositions and other belief-that-P-ish dispositions will be tenuous to the extent these three conditions are not met. (The first two of these conditions resemble the conditions of the "SOKA" model of self-knowledge of personality advocated in Vazire 2010.) When self-ascription is evaluatively neutral, self-deception and self-flattery are beside the point; when the connection to behavior is straightforward, self-deception and self-flattery have less room to play; and when the pertinent dispositions mostly concern judgment or utterance of P, transparency strategies for self-ascription are likely to be successful.

Consider the belief that Sacramento is the capital of California, as held by a schoolchild in Mississippi. All three conditions are fairly well met: Believing that proposition reflects

neither well nor poorly on the child except insofar as the proposition is true and it looks good to believe true things. Belief in this proposition is straightforwardly connected to observable behavior, namely, answering certain questions on exams, drawing a star next to the city on a state map, reciting lists of state capitals. And the dispositions most relevant to the ascription are avowals or judgments or statements that P is the case. What we tend to care about most in assessing whether the child has the belief is what she will say the capital is. Thus, my model predicts what seems very plausibly to be the case: Mississippi schoolchildren know fairly well whether they believe that Sacramento is California's capital.

Note that the third condition – concerning what dispositions are most relevant to the ascription – is target-and-context sensitive. If the subject is not a Mississippi schoolchild but an adult citizen of California, the dispositions that we might consider important in the context of ascription fan out more broadly, including where she thinks the legislature convenes, where she assumes she will be sending certain forms, where she thinks her son's debate team will be travelling when she hears they will be going to the capital for finals, etc. If enough of these other dispositions are out of line, we might want to describe her as mixed-up or in-betweenish regarding the location of the capital, even if she would say, when asked directly, both that the capital is Sacramento and that she thinks or believes or knows that it is Sacramento.

At the other end of the spectrum, consider the belief that those below you in social status deserve equal respect. Possession of this belief is normatively loaded: It seems not only mistaken but also rather arrogant and elitist not to believe it. Its connection with observable behavior is not entirely straightforward. There's enough of a gap between behavior and possession of the attitude that someone motivated to see the subject a certain way – including of course the subject himself – can easily rationalize away apparent counter-tendencies. Piotr, let's

say, has been rather short-tempered with the cashier. He might not reflect on the relationship between this fact and his disposition to avow equal respect for all, but if he does reflect, it's easy to concoct rationalizations to explain away the unappealing conclusion: Piotr might say that he is equally short-tempered with everyone, or that this particular cashier deserved it, or that he was unusually anxious about work at the time – any of which may or may not be true. If the schoolchild says "San Francisco" when asked the capital of California, the available excusers, while not non-existent, are not quite so happily manifold. And finally, a broad range of dispositions are relevant to the ascription of belief that those below you in social status deserve equal respect. While we *might* mostly care about what Piotr would say or explicitly judge, as likely as not we care just as much about his broad range of behavior – especially, perhaps, if we are among his maltreated inferiors. And clearly Piotr's self-knowledge of his mixed-up, in-betweenish condition might not be very good.

Let's flesh out the case. Maybe Piotr – the Elmo P. Schnerbuckle Professor of History and Classics at an Ivy League university – will sincerely, unhesitantly, and unqualifiedly assert, with a feeling of inward assent, that all people deserve equal respect, including those below him in social status. Piotr's possession of that disposition is perfectly consistent with the following facts about him: He tends to be disrespectful of people below him in social status, and to be so in a way that he does not generally, on reflection, find inappropriate; he is much more respectful, in a way he finds fitting upon reflection, of those above him in social status; and he also judges it fitting when he sees other people hierarchically distribute their acts and tokens of respect. So, for example, Piotr is not inclined to see much wrong with treating a housekeeper as a general-purpose servant – as long as he himself or some other high-status person is doing so. (If a lower-status person were to do so, say, some low-class tourist – he would find it jarringly

inappropriate.) It only seems right to Piotr that high-status people, himself and others, should not have to wait in line at the post office with all the schmoes; he finds the egalitarianism of the post-office line rather annoying – though he might not be quite willing to say this directly, either aloud or to himself. If Piotr were to see the president of the university in a middle seat of an airplane's coach section, that would strike him as a situation calling for rectification. He also implicitly expects the flight attendant to discern the president's high status and consequently to accord him more attention and respect. When the flight attendant treats the university president, and Piotr himself, in the same brusque manner as she treats the other coach passengers, Piotr feels she has acted somewhat inappropriately.

Piotr may or may not know these facts about himself; I see no reason to think he would have any specially privileged self-knowledge about such matters, compared to other people who have seen enough of Piotr's relevant behavior. He may well know what he would *say* about the proposition that all people deserve equal respect, how he would judge the proposition considered explicitly. But to the extent that believing a proposition is not only a matter of being disposed to sincerely assert but also a matter of how one steers one's way through the world – a matter of how one acts and reacts in a wide variety of situations, what one implicitly assumes, what is exhibited in one's reasoning and emotional reactions and spontaneous behavior – to that extent, any privilege that attaches to one's knowledge of what one would say or explicitly judge extends only contingently to the matter of what one believes in toto. If my sincere avowals don't align with my overall dispositional structure regarding P, then too heavy a reliance on transparency strategies, or on other simple and seemingly privileged strategies for self-ascription, will lead me astray. I might think I straightforwardly believe, when in fact my dispositional structure is a mixed-up muddle. On the other hand, we aren't always so naive about ourselves: Piotr could

well recall his relevant behavior when he reflects on what he believes, or he could imagine his reactions to certain situations – and then he might give a more nuanced self-ascription, or at least hesitate. And even if Piotr does neither, it doesn't follow that self-interpretive strategies, memories of his own behavior, and elitist impulses aren't feeding into the process issuing in self-ascription, even if those contributing processes are swamped or trumped.

5. Wraiths of Judgment.

Does Piotr know, at least, what he is currently judging or avowing, even if he doesn't know so well what he dispositionally believes? Well, what is it, anyway, to judge that all people deserve equal respect? It's an appealing form of words. But does the judgment include, as a facet, that the university president should have no better a chance for an aisle seat than anyone else? Or that eminent professors shouldn't treat housekeepers as general-purpose servants? I suppose that's not entirely clear. The unclarities here are part of what allow Piotr to retain his self-flattering self-attribution of egalitarianism in the face of the evidence. He will tend to interpret the unclarities in a certain way – which itself is a manifestation of his inegalitarian dispositions. Could a judgment like "all people deserve equal respect" be so wholly thin and abstract that Piotr could reach it despite rejecting *all* specific manifestations of egalitarianism?

The feature of the case I want to highlight is best displayed by shifting to a more extreme example of unclarity. Most citizens of the United States will endorse the familiar proposition that "all men are created equal" from the Declaration of Independence. How can you deny *that?* But what does it mean, really? And what would be involved in living your life with that among the guiding truths? It is a form of words we have learned to endorse, that has an appealing ring, that we associate with good things like democracy, but when we say it, however sincerely, do we

really judge it to be the case? I'm not sure there is any particular proposition we are judging to be true when we say that uplifting-sounding thing. Perhaps here we can compare ourselves with Adolf Eichmann, in Hannah Arendt's (1963) portrayal of him: We are all to some extent drawn to avow thrilling, evocative sentences that are only half-filled or quarter-filled with a real thought. If so, we might think that we are judging that all men are created equal, or that human life has infinite value, or that God died for our sins, or that the only really important thing is to be happy, or that Thousand Oaks High School is the greatest, when in reality we are only echoing pleasing phrases inhabited by merely the wraith of a thought, phrases possibly expressing an emotional reaction more than a judged or believed proposition. If we think we believe such things as we say them, or if we think that we even momentarily judge them to be true, we might quite well be wrong. I can say "damn it!" without hoping that something will be damned; I can say I'm all for freedom and brotherhood, embracing the form of the words and unbeknownst to me very little of their substance.

Similarly, it seems possible for abstract or abstruse philosophical views to earn sincere endorsement with only half- or quarter-full judgment. A student who is only starting to get an inkling of Kant might think that she is really judging, along with Kant, that "pure a priori concepts, if such exist, cannot indeed contain anything empirical; yet, none the less, they can serve solely as a priori conditions of a possible experience". She might say those words in inner speech and feel some sort of assent, but not really reach a judgment with that as the propositional content. This is perhaps especially clear if judging that P is partly defined in terms of having the right sorts of functional connections to other P-related thoughts and behaviors.

It also seems possible to self-attribute or assert distractedly – to absent-mindedly, as it were, say "yeah, I think you're right" or "yeah, what a loser" – where this is almost just knee-

jerk speech, with again only the shadow of a real judgment invested in it. In such cases, if you then pull up and reflect about what you just said – a moment later, but perhaps still within the "specious present" – you may or may not have an accurate idea of the extent to which the speech reflected a genuine judgment.

Somewhat differently: Walking to class, I remind myself what a privilege and honor it is to be permitted to teach 500 students. When a valued colleague accepts an offer to teach at another university, I may say to myself that there is nothing wrong with her doing so. When an editor returns my journal article with a slew of suggestions for revision, I assure myself that this is much better than having the warty thing accepted as-is. Piotr might, in the same spirit, remind himself that those below him in status deserve equal respect. Am I convinced of these things I tell myself – or is it more that I am *trying* to convince myself? And if I am trying to convince myself, how successful have I been? To what extent do I really judge these things to be true, as I say them? I'm not really sure. Here again, privilege fails.

Such indeterminacy about whether I am reaching or articulating a judgment can also affect the self-ascriptive act itself. It seems to me that most philosophical accounts of self-knowledge try to divide affairs too sharply on this matter. Sometimes, when I say something like "I think it's going to rain" I am not at all reaching or articulating a judgment about what I think or believe or judge. It is *simple assertion* that happens to use self-ascriptive language, without any accompanying self-ascriptive cognition. (Perhaps the absence of self-ascriptive cognition is clearest with two-year-olds who can apparently use a sentence like "wanna cookie" to mean "gimme cookie" without any self-ascriptive metacognition at all; Gordon 2007.) At other times, I can use the very same words "I think it's going to rain" as an *avowal* – that is, as a speech act with two simultaneous aims, one of those aims being self-ascriptive: I am articulating or

conveying a judgment both about the weather and simultaneously about my mind. (Rarely, there might also be cases where I really am only articulating a judgment about my own mind.) I suspect, however, that many cases – perhaps most cases – aren't cleanly one or the other but somewhere between. That is, I suspect that in many cases I am somewhere between making a simple assertion that articulates a judgment only about the weather and making an avowal that articulates a judgment *both* about the weather and about my mind. Such intermediate cases might involve, as it were, a hint or shadow of a judgment about my mind alongside my primary judgment about the weather, a bare smidgen of the phenomenology and/or functional role constitutive of the self-ascriptive judgment. Similarly, when someone asks me whether I think there will be a third world war, if I use a transparency strategy to reply, my strategy might be somewhere between the topic-shifting and self-judgment versions of transparency, or an amalgam of the two strategies.

The same holds of utterances that are not literally self-ascriptive: Sometimes when I say P I have the self-expressive intent of dual-purpose avowal. I mean to convey something both about the world and about my mind. Other times, my own mind is, as it were, the further thing from my mind. But often the case might be intermediate between these two extremes. Why must judgments and utterances all neatly partition into ones that determinately involve self-ascription and those that determinately do not?

6. Conclusion.

Thus, I invite you conceptualize the self-knowledge of belief as follows. To believe is to possess a wide variety of dispositions pertinent to the proposition believed. Among the pertinent dispositions are self-attributive dispositions. Consequently, being disposed to self-attribute

belief that P, or more generally believing that one believes that P, is partly constitutive of believing that P. These self-attributive dispositions can be underwritten by any of a variety of mechanisms, acting co-operatively or competitively. But since the self-attributive dispositions are only *partly* constitutive of belief, there can be cases in which the self-attributive dispositions splinter away from the remaining dispositions. In such cases, we will be prone to self-attribute belief when in fact we don't believe or when in fact our dispositional profile is muddy and in-betweenish. It is then an empirical question how often our self-attributive dispositions diverge from the remaining dispositions constitutive of belief. The dispositions will tend to align reliably when possession of the belief in question is normatively neutral, straightforwardly connected to behavior, and most centrally (given the purposes of the ascriber) manifested in explicit assertions or judgments. When these three conditions are not met, self-ascription, self-conscious avowal, sincere or for-all-the-subject-can-tell sincere utterance, and explicit judgment will often diverge from the various other dispositions constitutive of belief. Even self-knowledge of what we explicitly judge on some particular occasion (in contrast to what we dispositionally believe as a general matter) can be problematic when we are attracted to half-empty forms of words and when we are not, or may not be, entirely persuaded of what we are saying. Nor need it always be a determinate matter whether we are reaching a self-ascriptive judgment or not.

References.

Arendt, Hannah (1963). *Eichmann in Jerusalem.* New York: Penguin.

Armstrong, David M. (1968). *A materialist theory of mind.* New York: Routledge.

Armstrong, David M. (1981). *The nature of mind and other essays.* Ithaca, NY: Cornell.

Armstrong, David M. (1999). *The mind-body problem.* Boulder, CO: Westview.

Bar-On, Dorit (2004). *Speaking my mind.* Oxford: Oxford.

Bartlett, Frederic C. (1932). *Remembering.* Cambridge: Cambridge.

Bem, Daryl J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review, 74,* 183-200.

Braithwaite, R.B. (1932-1933). The nature of believing. *Proceedings of the Aristotelian Society, 33,* 129-146.

Byrne, Alex (2005). Introspection. *Philosophical Topics, 33* (1), 79-104.

Carruthers, Peter (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences, 32,* 121-182.

Carruthers, Peter (forthcoming). *The opacity of mind.* Oxford.

Cartwright, Nancy (1983). *How the laws of physics lie.* Oxford: Oxford.

Dretske, Fred (1995). *Naturalizing the mind.* Cambridge, MA: MIT.

Evans, Gareth (1982). *Varieties of reference.* Oxford: Oxford.

Goldman, Alvin I. (2006). *Simulating minds.* Oxford: Oxford.

Gopnik, Alison (1993a). How we know our own minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences, 16,* 1-14.

Gopnik, Alison (1993b). Psychopsychology. *Consciousness and Cognition, 2,* 264-280.

Gordon, Robert M. (2007). Ascent routines for propositional attitudes. *Synthese, 159,* 151-165.

Heil. John (1988). Privileged access. *Mind, 97,* 238-251.

Hunter, David (forthcoming). Alienated belief. *Dialectica.*

Hurlburt, Russell T., and Eric Schwitzgebel (2011). Presuppositions and background assumptions. *Journal of Consciousness Studies, 18* (1), 206-233.

Kripke, Saul (1979). A puzzle about belief. In A. Margalit, ed., *Meaning and use.* Dordrecht: Reidel.

Lawlor, Krista (2008). Knowing beliefs, seeking causes. *American Imago, 65,* 335-356.

Lewis, David (1980). Mad pain and Martian pain. In N. Block, ed., *Readings in philosophy of psychology, vol. 1.* Cambridge, MA: Harvard.

Lycan, William G. (1996). *Consciousness and experience.* Cambridge, MA: MIT.

Marcus, Ruth B. (1990). Some revisionary proposals about belief and believing. *Philosophy and Phenomenological Research, 50,* 132-153.

McGeer, Victoria (2007a). The moral development of first-person authority. *European Journal of Philosophy, 16,* 81-108.

McGeer, Victoria (2007b). The regulative dimension of folk psychology. In D. Hutto and M. Ratcliffe, eds., *Folk psychology re-assessed.* Dordrecht: Springer.

McGeer, Victoria, and Philip Pettit (2002). The self-regulating mind. *Language and Communication, 22,* 281-299.

Moran, Richard (2001). *Authority and estrangement.* Princeton: Princeton.

Nichols, Shaun, and Stephen P. Stich (2003). *Mindreading.* Oxford: Oxford.

Price, H.H. (1969). Belief. London: George Allen & Unwin.

Roediger, Henry L., III (1980). Memory metaphors in cognitive psychology. *Memory and Cognition, 8,* 234-246.

Ryle, Gilbert (1949). *The concept of mind.* New York: Barnes & Noble.

Schwitzgebel, Eric (1999). Gradual belief change in children. *Human Development, 42,* 283-296.

Schwitzgebel, Eric (2001). In-between believing. *Philosophical Quarterly, 51,* 76-82.

Schwitzgebel, Eric (2002). A phenomenal, dispositional account of belief. *Noûs, 36,* 249-275.

Schwitzgebel, Eric (2006/2010). Belief. *Stanford Encyclopedia of Philosophy.* URL: http://plato.stanford.edu/entries/belief

Schwitzgebel, Eric (2010). Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly, 91,* 531-553.

Schwitzgebel, Eric (forthcoming). Mad belief? *NeuroEthics.*

Shah, Nishi, and J. David Velleman (2005). Doxastic deliberation. *Philosophical Review, 114,* 497-534.

Shoemaker, Sydney (1996). *The first-person perspective and other essays.* Cambridge: Cambridge.

Shoemaker, Sydney (2009). Self-intimation and second order belief. *Erkenntnis, 71,* 35-51.

Sutton, John (1998). *Philosophy and memory traces.* Cambridge: Cambridge.

Tye, Michael (2000). *Consciousness, color, and content.* Cambridge, MA: MIT.

Vazire, Simine (2010). Who knows what about a person? The Self-Other Knowledge Asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 98,* 281-300.