Philos Stud (2019) 176:559–579
https://doi.org/10.1007/s11098-017-1028-9

Introspecting knowledge

John Gibbons[1]

Published online: 17 January 2018

Abstract If we use "introspection" just as a label for that essentially first-person way we have of knowing about our own mental states, then it's pretty obvious that if there is such a thing as introspection, we know on that basis what we believe, and want, and intend, at least in many ordinary cases. I assume there is such a thing as introspection. So I think the hard question is how it works. But can you know that you know on the basis of introspection? Well, that all depends on how introspection works. I present one account of how introspection works and argue that on that account, you can know that you know ordinary empirical things on the basis of introspection. As far as how we know about them is concerned, there's no principled difference between the factive and non-factive mental states.

Keywords Introspection · Knowledge · Self-knowledge · Transparency

Sometimes you know that the door is open. And sometimes you know that you know. This seems like a safe place to start. Broadly speaking, there are two routes forward from here. On the one hand, we can ask, in as naïve a philosophical voice as we can muster, "How does knowing that you know work?" On the other hand, we might leap immediately to some kind of general principle. Perhaps this comes to mind.

$$(\mathrm{KK})\ \Box\,(p)\,[\mathrm{K}p \rightarrow \mathrm{KK}p]$$

✉ John Gibbons
john.gibbons@st-hildas.ox.ac.uk

[1] St Hilda's College, University of Oxford, Cowley Place, Oxford OX4 1DY, UK

Necessarily, for all p, if you know that p, then you know that you know that p. Given the obvious regress, I take it that this is out of the question. The KK Thesis is less of a view than a trap you might fall into if you're not careful.

Perhaps we should start down the other road. How does knowing that you know work? One of the most striking things about knowing that you know is the lack of anything that looks obviously like a method, or a procedure, or a way of finding out. How do you know if the door is open? You look. It's not an elaborate procedure, but it's something you can do. How do you know that you know the door is open? Looking again doesn't help.

So maybe what's behind the search for general principles is not just the love of generality. Maybe the idea is that there just isn't all that much to second-order knowledge over and above first-order knowledge. And maybe, the thought continues, if there's a logical connection between the facts, the fact that you know that p and the fact that you know that you know that p, this might explain the apparent lack of a procedure and the apparent ease with which second-order knowledge comes when it does. But even if we give up on the general principles, it is worth noting the apparent ease and apparent lack of a procedure when it comes to knowing that you know. And perhaps we should add that in a surprising number of cases, you are in a better position to know that you know than I am to know that you know. How's that work?

Sometimes you know what you're thinking. This seems like a safe place to start. But it's easy to get overly enthusiastic about self-knowledge. We're all familiar with exaggerated philosophical claims about the scope, depth, reliability, and authority of our own views about our own minds. And we're also familiar with the idea that any distinctively first-personal knowledge we have of our own mental states must be restricted to the inner life.

Both the enthusiasm and the restriction can be motivated by thinking about a certain kind of skeptical scenario. Suppose you're a brain in a vat. You could be wrong about all sorts of things. But the mere fact that you're a brain in a vat doesn't call into question your knowledge of your own beliefs, desires, and experiences. So maybe what's distinctive about our knowledge of our own mental states is its immunity from a certain kind of skeptical scenario. And maybe this immunity warrants some enthusiasm.

But what starts off as an interesting epistemological discovery, the idea that our knowledge of some mental states is immune to the relevant scenarios, threatens to turn into a definition of the mental. Knowledge, perception, and intentional action are the kinds of thing that wouldn't happen in a world without minds. So they're at least a little mental. But if you were a brain in a vat, you could easily be wrong in thinking that you see a cat. You could be wrong about what you know and what you're doing on purpose even if you weren't wrong about how things seem, what you believe, and what you're trying to do.

So maybe knowledge, perception, and intentional action aren't really mental, or anyway, not purely mental. And they're not mental because our knowledge of them is not immune to the relevant scenarios. But now our interesting epistemological discovery has turned into the idea that anything immune to the scenarios is immune to the scenarios. It's not clear how much enthusiasm this warrants. But it does make

sense of lumping our knowledge of our own inner life in with our knowledge of easy math and basic logic, no matter how metaphysically and epistemologically distinct those things may otherwise have seemed.

Immunity to a certain kind of skeptical scenario is one thing. Infallibility is quite another. But maybe what's behind the temptation for infallibility is the idea is that there just isn't all that much to knowing that you believe that *p* over and above believing that *p*. And maybe, the thought continues, if there's a logical connection between the facts, the fact that you believe that *p* and the fact that you know you believe that *p*, this might explain the apparent lack of a procedure and the apparent ease with which self-knowledge comes when it does.

The problem with the infallibilist's position, of course, is that it's just so implausible. For some of us, self-deception is a daily occurrence. But in our zeal to reform the Over-Enthusiasts, we shouldn't forget just how easy self-knowledge is. Usually, there's not that much you have to do to figure out whether you believe that *p*. And perhaps we should add that in a surprising number of cases, you are in a better position to know what you believe than I am to know what you believe. How's that work?

There's another way of thinking about essentially first-personal knowledge that's not based on the idea of immunity to error. On the one hand, there's the idea of a distinctive route. You have a way knowing about your own mental states that depends essentially on the fact that they're yours. What's thought and thought about depend on the same point of view. So this is one part of the first-person/third-person asymmetry. My way of forming beliefs about your mental states is different from yours. Maybe it's less direct or immediate.

And on the other hand, there's the idea of a special status. The route, whatever it is, confers a high degree of justification, or warrant, or entitlement, or whatever you like. This is another part of the asymmetry. Your beliefs about your own mind have better epistemic credentials than my beliefs about your mind, at least in the ordinary, everyday case.

So we start with the idea that when we're talking about self-knowledge, we're not just talking about any old knowledge you have about yourself. We're talking about your knowledge of your own mental states. But we shouldn't be absolutely certain that only facts of a specific metaphysical kind are available to us in a first-person way: only facts that are inner and present can be known this way.

In the ordinary, everyday case, if you already have a plan, then you know what you'll be having for dinner. And the fact that makes your belief true is not only all the way out there in your kitchen. It's all the way out there in the future. But if you know what's going to happen because you've made up your mind that it's going to happen, then your knowledge of what you will be doing might be essentially first-personal after all (Anscombe 1957; Gibbons 2010). If what we're after is something we couldn't possibly be wrong about no matter what, we should abandon all hope. But if what we're after is something essentially first personal, then maybe your knowledge of your intentional actions could be like that. We should wait and see.

If we give up on infallibility and the idea that there's a logical connection between first-and second-order belief, there's no longer an obvious motivation for the retreat to the inner. Typically, often enough, and with surprising ease, we know

what we want, and we know what we believe. And we should add to this list of standard examples. Typically, often enough, and with surprising ease, we know that we know. We know what we can see from where we are. And we know what we're doing on purpose. Our way of knowing these things about ourselves is very much like other people's ways of knowing the same things about themselves. And it's not that much like other people's way of knowing the same things about us. There's something essentially first personal about this way of knowing.

I think that knowing that you know is an example of plain old regular self-knowledge. It's not the kind of thing you need to go to a therapist for. In the first part of this paper, I'll sketch a picture of how ordinary self-knowledge works and look at some objections to extending the picture to the case of knowing that you know. In the second part, I'll raise what I take to be the fundamental question about the picture and sketch an answer to that question. To the extent that the picture and the answer work, they work just as well for knowing that you know as they do for knowing that you believe.

1 Part I: the picture

1.1 The route and the status

There are two interesting things about plain old regular self-knowledge. On the one hand, there's the idea of a distinctive route. My way of forming beliefs about your mental states is different from yours. And on the other hand, there's the idea of a special status. Your beliefs about your own mind typically have better epistemic credentials than my beliefs about your mind. This raises three questions: What's the route? What's the status? And why does the route confer the status?

I think of the distinctive route in terms of the idea of transparency. Many people like this idea, and different philosophers give different accounts of what it comes to (Edgley 1969; Evans 1982; Moran 2001). Eventually, I'll give you my account of what it comes to. But for now, we start with the basic idea.

> (Transparency) It's okay, and probably more than just okay, to answer questions about the mind by thinking about the world.

You answer the question of whether you believe that *p* by thinking about whether or not *p*. But you also answer questions about what you want for dinner by thinking about food. At least, I do. But the idea is not just that we do it this way. The idea is that it's epistemically okay to do it this way. And this can seem kind of puzzling. Why is it okay to answer questions about one thing by thinking about something you know to be different? But some people think it's more than just okay (Moran 2001). Even if you did it some other way, if you couldn't do it this way, that would show that there's something seriously wrong with you in the rationality department. And this can seem even more puzzling. We'll come back to transparency.

What's the status? In the olden days, they used to think about this in terms of some sort of general principle: infallibility, indubitability, incorrigibility, etc.

(Alston 1971). I assume we're all past this by now, or anyway, think we are. But infallibility and friends are just special cases of a more general strategy: counting kinds of possibility for error.

So the standard view is that we have privileged access to our non-factive mental states like beliefs, desires, and experiences. But we don't have privileged access to our factive mental states and events like knowledge, perception, and intentional action. Here's one way of making that view seem plausible. Suppose you have a false belief that p. You may well correctly believe that you believe that p, but if you believe that you know that p, that would be a mistake. Now suppose you believe that you believe that p because you desperately want p to be true. You don't really believe that p. Now your belief that you believe that p is false, and so is your belief that you know that p.

You're wrong about whether you believe only in the second case, but you're wrong about whether you know in both cases. So there are more kinds of error possibilities for first-person knowledge claims than there are for belief claims. So either we have more privileged access to our beliefs than to our knowledge, or, as the standard view would have it, we have privileged access to what we believe, but not to what we know.

But if you're counting kinds of error possibilities, then the best-case scenario is zero possibilities for error. And that's just what the traditional definition of infallibility says.

(Infallibility) $\Box\,(p \in \mathrm{R})\,(\mathrm{B}p \rightarrow p)$

Necessarily, for every proposition within a certain range, if you believe it then it's true. Now let R be the set of necessary truths. If you believe any one of them, you're guaranteed to be right. But this has no epistemic consequences whatsoever. Most importantly, it says nothing about why you believe that p. If you believe that arithmetic is incomplete because it just seems that way, you're guaranteed to be right. But this says nothing about the epistemic credentials of the belief because the epistemic credentials of a belief depend on what it's based on.

1.2 When justifiers are also truth makers

There's another way of thinking about the special status. Suppose that your belief that p is the reason for which you believe that you believe that p. I represent that like this.

$$\frac{\mathrm{B}p}{\mathrm{BB}_p}\ \mathrm{RFW}$$

This is not an argument. This is a description of a transition that may or may not occur. But if it does occur, the mental state on top is the reason for which you go into the state on the bottom. This is not an inference. If you infer that q from your beliefs that p and that if p then q, then the justification for the conclusion depends essentially on the justification for the premises. If your beliefs in the premises are not justified, then neither is your belief in the conclusion. But if those beliefs are justified, then so is your belief in the conclusion, and its justification is derived from

the justification of the beliefs it's based on. But you could have an unjustified, false belief that p and still know that you believe that p. So the justification of the second-order belief is not derived from the justification for the first-order belief.

We have a traditional model for this. Suppose it seems to you that p. On certain views, the experience itself justifies the belief that it seems to you that p. I represent that like this.

$$\frac{Sp}{\mathrm{BS}p}\ \mathbf{RFW}$$

And on certain views, its seeming to you that p, the experience itself, can justify you in believing that p, at least in the absence of reason to doubt.

$$\frac{Sp}{\mathrm{B}p}\ \mathbf{RFW}$$

Neither of these is an inference in the traditional sense. Assuming that experiences themselves are not justified, or warranted, or knowledge, the justification of the bottom state is not derived from that of the top. Inference transmits knowledge. It's not a source of knowledge. But introspection is a source of knowledge.

If we just look at the first two, the ones that involve a second-order belief on the bottom, I'm tempted to go on in the same way. The desire that p is itself your reason to believe that you desire that p.

$$\frac{Dp}{\mathrm{BD}p}\ \mathbf{RFW}$$

The idea that first-order states are reasons for second-order states about them appears in Peacocke (2000). But my view is dramatically less fancy than his. I think you can explain it all in terms of plain old regular rational causation. Ram Neta (2011) also presents a version of this view. We'll come back to him. So the basic idea isn't new. In fact, I've tried it out myself before (Gibbons 1996).

You can read this idea into Byrne (2005), if you understand rule following in terms of reasons for which. But Byrne does insist that it is an inference. It's just not the kind of inference that transmits knowledge, ignorance, or anything else. So Byrne would only accept my first example. According to him, the belief that o is desirable justifies your belief that you desire o (Byrne 2011). On Byrne's picture of rule following, the only thing left to the idea of inference is that it's got to be a belief on top.

I think my way of going on in the same way is better than Byrne's. In the language of reasons, the idea that it has to be a belief on top is the idea that the kinds of reasons that determine the rationality of the mental states they cause are always beliefs. So if you thought that the desire for the end, plus the belief that the means would achieve the end, can rationalize or make sense of taking the means, then you might, like Davidson (1963), call these things reasons for action. But if it's got to be a belief on top in order to be what Byrne calls inference, then either there's no such thing as practical inference, or desires are always and everywhere irrelevant to what it makes sense to do.

In epistemology, the idea that it has to be a belief on top can be expressed with Davidson's motto that "nothing can count as a reason for holding a belief except

another belief" (1986). On this view, experiences themselves are epistemically irrelevant, and you get coherentism. Coherentists tried to use beliefs about experience to play the epistemic role that experiences themselves actually play (BonJour 1985), and this was one of the main reasons for the downfall of coherentism (Sosa 1980). What does justify those beliefs about experience that are supposed to play experience's role? Randomly made up beliefs about experience are just not as good, epistemically speaking, as beliefs that are based on what they're about. So I'll go on in my own way.

Suppose that one of these transitions from the first order to the second order did occur. The first question to ask is if the resulting belief constitutes knowledge. Is it just an accident that your belief is true? No. The belief is justified on the basis of the fact that it's about. On one fairly common picture, the relation between justifiers and truth makers is usually not this close. According to the picture, in a case of perceptual knowledge, the fact that p causes it to seem to you that p, and its seeming to you that p justifies the belief that p. Assuming there's no funny business going on, the relation between the justifiers and the truth makers is close enough to constitute ordinary knowledge. But when the justifiers are themselves the truth makers, you get not only knowledge, but a special epistemic status.

Now suppose that this could happen, and it did.

$$\frac{Kp}{\mathrm{BK}p}\ \textbf{RFW}$$

Is it just an accident that your belief is true? No. In fact, given that the justifier is also the truth maker, it looks like privileged access.

Of course, this isn't what always happens. It only happens in the good case. But suppose you believe that you believe that p because you desperately want p to be true, but you don't really believe that p. You confuse one mental state for another. Surely, no one in this century would ever say out loud that this sort of thing could never happen. But this possibility isn't supposed to call into question the idea that in the ordinary, normal case, we have privileged access to our beliefs.

Now suppose you've got a false belief that p. This leads you to believe that you know that p. You've confused one metal state for another. Unless we simply assume the standard view, we have no reason to treat these two cases differently. But the standard view is exactly what's at issue. And if we don't treat the two cases differently, this possibility shouldn't call into question the idea that in the ordinary, normal case, we have privileged access to our knowledge.

1.3 Objections to extending

Ram Neta (2011) thinks there might be an objection to extending this picture to factive mental states. He doesn't quite argue for the standard view, but he thinks the best place to look for such an argument is in the following neighborhood. Maybe we can believe that we believe that p when and because we believe that p, but we can't believe that we know that p when and because we know that p. And we might think this is true because:

(A) You could easily know that p without believing that you know that p.
and (B) You could easily believe that you know that p but not know that p.

It's not clear what work (A) does. It is easy to know that p without believing that you know that p. But it's just as easy to believe that p without believing that you believe that p. As far as (A) is concerned, there's no difference between belief and knowledge. So if there's anything to this, it must be in (B).

So maybe the idea is that you could mistake a false belief for knowledge. But this is just confusing one mental state for another. If you say that this is impossible in the case of non-factive mental states but possible in the case of factive mental states, we'd have a principled difference between the factive and the non-factive. But I dare you to say that out loud.

So maybe the idea is that it's just easier to confuse one mental state for another in the factive case than in the non-factive case. You think you're in love, but it's really the other thing. How easy is that? Your tortures tell you that they're going to put their cigarette out on the back of your neck. But they put an ice cube there instead. You think it feels hot, but it really feels cold. Everybody makes this mistake. You think you're hungry, but you're really just bored, or upset, or stressed. There's nothing particularly weird about this, and it's really not the least bit uncommon.

If we think the line around our epistemic natural kind separates our knowledge of belief from our knowledge of knowledge, we need a principled difference between the non-factive and the factive. There is no principled difference.

1.4 Safety first

Over-Enthusiasts about self-knowledge are often chastised for underestimating how easy it is to get it wrong with respect to the inner life. But it's a different mistake to underestimate how easy it is to get it right when it comes to knowledge. It's plausible to suppose that safety is necessary for knowledge.[1] You know that p only if you couldn't easily be wrong in believing that p. Suppose you look at the door and come to believe that it's open. How easy is it for you to be wrong about that? It's not impossible. But there's a difference between *easy* and *possible*. And that's the difference between safety and skepticism. If this is an ordinary, everyday case, it's not that easy for you to be wrong about the door. The situations in which you're mistaken are not that close to the actual world. So in that kind of case, you could have ordinary empirical knowledge that the door is open.

So suppose you know that the door is open, and suppose further that you believe that you know. How easy is it for you to be mistaken in your second-order belief? It's not impossible. But are there *nearby* situations in which you're wrong in thinking that you know that p? They're not situations in which p is false. If there were nearby worlds where you think the door is open when it's not, then you don't know the door is open, at least if safety is necessary for knowledge. But we're assuming you do know the door is open.

[1] Sosa (1999), Pritchard (2005) and Williamson (2000). Humberstone (1992) contains an objection to the idea that safety is sufficient for knowledge.

The easiest kind of case to come up with where you're wrong in thinking that you know that *p* is the case where *p* is false. But first-order knowledge rules this case out not only as actual but also as nearby. So the second-order belief that you know is at least roughly as safe as the knowledge it's based on.

Gaining second-order knowledge from first-order knowledge is easy. It's not guaranteed, and it's not impossible for the two to come apart. If the route up from each order to the next were guaranteed, we'd get the familiar regress. You'd have to know infinitely many things in order to know anything. But if the route up from first-order to second-order belief were guaranteed, then you'd get a similar regress. You'd have to believe infinitely many things in order to believe anything. So the lack of a guarantee doesn't point to a principled difference between the factive and the non-factive.

1.5 Transmission failure

Here's one more objection to extending the picture to our knowledge of factive mental states. According to me, you can know that you know that *p* on the basis of introspection. And you could know a priori that knowing that *p* entails that *p* is true. If we define reflective knowledge as knowledge that's based only on introspection and the a priori, doesn't my theory entail that you could have reflective knowledge that the door is open? You just infer it from your introspective knowledge that you know plus your a priori knowledge of the entailment.

Whether or not an epistemic status will transfer over known entailment depends on what that status is (Wright 1985). Suppose that the status conferred by introspection is that you couldn't be wrong in believing that *p*. If your belief that *p* has this status, and you know that *p* entails *q*, then you couldn't be wrong in believing that *q*. In fact, this status would transfer over entailment whether known, or believed, or not. And this should make you wonder whether it's a distinctively epistemic status at all.

Suppose instead that the status conferred by introspection is that your belief that *p* is justified on the basis of the fact that it's about. Even if you know that *p* entails *q*, there's no reason to think that your belief that *q* will be justified on the basis of what it's about. That belief may well be based on your beliefs that *p* and that if *p* then *q*.

If we think of the special status in terms of what's based on what, we should not expect it to transfer over known entailment. And when you look at the picture of how things go when things go well, we should expect that it would not transfer in these cases. If I'm right, your introspective knowledge that you know is based on your knowledge that the door is open, and not the other way around. So the first-order belief is not based on introspection and the a priori.

2 Part II: the question

So there's a picture of how self-knowledge works when it does, and the picture applies equally to the factive and non-factive mental states. And the picture tells you what the special status is: it's when justifiers are also truth makers. When the belief

is justified on the basis of the fact that it's about, you get not only knowledge but a special status.

And the picture explains why the route confers the status. If your therapist convinces you that deep down, you believe your mom is out to get you, you have a belief about your beliefs. But this is not the kind of self-knowledge we're talking about. And this second-order belief is not based on what it's about. It comes from trusting your therapist. And the picture explains the first-person/third-person asymmetry. Your mental states can directly cause your beliefs about them, but they can't directly cause my beliefs about them. And that's because what's thought and thought about are in the same mind.

So far, so good. Let's look again at our transitions.

$$\frac{Bp}{BBp}\ \textbf{RFW} \qquad \frac{Sp}{BSp}\ \textbf{RFW} \qquad \frac{Dp}{BDp}\ \textbf{RFW} \qquad \frac{Kp}{BKp}\ \textbf{RFW}$$

Aren't they lovely? There are various things to be said in favor of these transitions. They're reliable. In fact, they're hyper-reliable. If you form the second-order belief in this way, that belief is guaranteed to be true. Furthermore, we philosophers can know a priori that the transitions are hyper-reliable. That's pretty good. But suppose someone believes that the sky is blue, and concludes on that basis that arithmetic is incomplete. We philosophers can know a priori that this transition is hyper-reliable. But we're not impressed with the epistemic credentials of this belief. Thinking about what's good about the transitions in terms of a guarantee of truth is just the same thing as thinking about the special status in terms of counting kinds of error possibilities. And this explanation leaves many of us feeling a little cold.

But we all have an inner tortoise (Carroll 1895). We know that we shouldn't listen to our inner tortoise. But it's still in there. What our inner tortoise wants is not just that the transition is hyper-reliable or that somebody else knows that it's hyper-reliable. What your inner tortoise really wants is that the subject knows that the transition is hyper-reliable. The tortoise wants the subject to see the connection between the thing on top and the thing on the bottom. And seeing the connection must be a third mental state distinct from the things on top and the bottom. It makes no difference whether you call the third mental state accepting a premise or accepting a rule. It's a mental state whose epistemic credentials can be called into question.

Suppose you arrive at the belief that q on the basis of inference from something else. And suppose it looks at first as though there are twelve necessary conditions for your belief that q to be justified. If you're not aware of one of these conditions, then it is for you as if it's not there. But it's a necessary condition on justification. So it is for you as if you're not justified. So your belief that q is not responsibly formed. So you need to be aware of them all. But that means that the awareness of the first twelve necessary conditions is itself a necessary condition on justification, and you need to be aware of it. If you like, you can say that you only need to be aware of the first three of the first twelve necessary conditions.[2] Now you have conditions on

[2] See Leite (2008) for a view of this sort.

justification it is logically possible to satisfy, but you won't satisfy your inner tortoise.

We don't really trust our inner tortoise. So even if we're left feeling a little cold by hyper-reliability, we're also left wondering if we should be. If the only options are the tortoise's impossible conditions and hyper-reliability, then maybe hyper-reliability is the best that we can do.

2.1 When questions are connected

But even if we ignore our inner tortoise, a serious question about our transitions remains. Why is the thing on top a reason for the thing on the bottom? This is the fundamental question about the picture. One problem with moving from the sky's being blue to the incompleteness of arithmetic is that the premise is about one thing, and the conclusion is about another. And something similar is going on with our transitions. So when we ask why the thing on top is a reason for the thing on the bottom, here's one thing that might be bothering us. How could it be okay to answer questions about one thing by thinking about something you know to be different? And this brings us back to transparency.

Suppose someone asks you in a non-philosophical context if you know where the keys are. You know, but don't always notice, that this is a question about you. In the good case, you say, "Yeah," thereby answering the question of whether or not you know. But unless you've trained yourself to be annoying, you immediately follow this with, "They're on the dining room table." You don't always notice that this is a question about you because you answer the question about yourself by thinking about the world. There's no need to force knowing that you know into the mold of introspection. It's already in the mold.

Yeah, but how's that work? It's okay to answer questions about one thing by thinking about something else when the questions are connected. And questions can be connected in different ways. Here's one kind of case where the questions are connected by way of a mental state that connects them. The questions are connected because you believe that the facts are connected. Suppose you know that if p then q, but you don't yet have a view about either p or q. Now the question of p is not independent of the question of q. And they're connected by way of your knowledge of the conditional. So you have to answer these questions together.

Some cases are messier than our original case. Here you have no view about whether or not if p then q; no view about p; and no view about q. But the questions have come up, and you have to answer them together. Any evidence that p needs to be weighed against any evidence you might have for (if p then q) and not-q. And so on for all the relevant combinations.

Why are these three questions connected? Is it by way of a linking mental state, perhaps your antecedent acceptance of the tortoise's first premise?

(T1) $((p\&(p \rightarrow q)) \rightarrow q)$

But what gives you the right to believe that? And calling it a rule changes nothing except how we ask the question. What gives you the right to accept the rule? Is it

because you have to know logic without reasoning before you can reason? Or is it that really, there are four connected questions?

$$p? \quad (p \to q)? \quad (\text{T1})? \quad \text{and} \quad q?$$

And you need the tortoise's second premise, or the corresponding rule, to connect them. It's a very difficult question exactly how all this works. But one thing we know for sure is that the questions don't always have to be connected by way of a linking mental state.

If you know that if p then q, you see the connection between p and q. If you also believe p, you need to put these two mental states together to get anything out of them. But putting them together approximately amounts to thinking about them at the same time. You don't need anything other than this, aside from being reasonable, to get something out of them. And being reasonable is not a third mental state whose epistemic credentials can be called into question. It's not a matter of wanting to be reasonable; it's not a matter of having a second-order intention to comply with the demands of reason whatever those may be; and it's not a matter of knowing all of math and logic. It's a matter of responding to good reasons because they're good reasons.

And this gives us some clue about what good reasons are. They're the kind of thing that can directly get you to avoid the bad combinations. Sometimes you revise. Sometimes you conclude. But the reasons themselves can keep you from believing p; believing if p then q; and believing not-q, at least when you're being reasonable. But good reasons don't just get you to avoid the bad. They can also get you to believe what you ought to believe and justify what they cause.

It's tempting to represent certain sorts of inferences in the following way.

$$(\text{MoPo}) \quad \frac{\mathrm{B}p \qquad \mathrm{B}(p \to q)}{\mathrm{B}q}\ \textbf{RFW}$$

In this case, the justification for your belief that q is determined by the justification for your beliefs that p and that if p then q. And what makes this transition rational does the same thing for the transitions you make both in the original case and the messy case of answering questions together.

What does make these transitions rational? One thing that ties these three together can be understood in terms of a wide-scope "ought" (Broome 1999). Suppose the questions of p, q, and if p then q have come up and they matter to you. If you do answer all these questions, the worst-case scenario this. You believe that p; you believe that if p then q; and you believe that not-q. But the second-worst-case scenario is not that great. Here, you believe that p; you believe that if p then q; you're trying to figure out whether or not q; and you have no clue.

Here's one way of putting the relevant wide-scope "ought" in English. If you're going to have views on these matters, your views ought to be like this: You believe that p and that if p then q only if you believe that q. Here's how that looks in shorthand.

$$(\text{WSO}) \qquad \mathrm{O}\,[(\mathrm{B}p \,\&\, \mathrm{B}(p \to q)) \to \mathrm{B}q]$$

You can satisfy this requirement of reason either by believing that *q*, or by failing to believe either *p* or if *p* then *q*.

In the first instance, the wide-scope "ought" rules out the bad combinations. In effect, it tells you what to avoid. Don't believe this stuff unless you also believe that stuff. More than that, it doesn't on its own tell you what to do. But while (WSO) doesn't explicitly say anything about transitions like (MoPo), we can use it to explain why the transitions are rational.

We assume that the questions of *p* and so on have come up, and we think of (WSO) as ruling out as impermissible certain combinations of answers. From the mere fact that you do believe that *p* and that if *p* then *q*, the main thing that follows is that you ought to either revise or conclude. But now suppose that you ought to believe that *p*, and you ought to believe that if *p* then *q*. When it comes to the requirements of rationality, I think that what you ought to believe is determined by the evidence, broadly construed.

(WSO), on its own, says that you ought to either revise or conclude. Your evidence for *p* and for if *p* then *q*, says that you shouldn't revise. So your only permissible option left is to conclude. So that's what you ought to do. In this sort of case, a transition like (MoPo) could be rationally required. Given your evidence, it's the only reasonable move for the mind to make. If the evidence had been otherwise, (WSO) would license the transition from the beliefs that if *p* then *q* and that not-*q* to the belief that not-*p*. At least sometimes, there's only one way to go to avoid the bad combinations.

You can think of (WSO) as the idea that you are epistemically responsible for some, though certainly not all, of the logical facts themselves. You're responsible for the obvious ones. (WSO) doesn't just rule out the worst-case scenario where you believe that *p*, believe that if *p* then *q*, and believe that not-*q*. It also rules out the second-worst-case scenario where you believe that *p*, believe that if *p* then *q*, and have no clue about *q*. If the question has come up and you can't figure out the answer, this is a failure of rationality.

If true, (WSO) is true regardless of whether you believe it, accept it, or represent it. In the language of reasons, if your justified beliefs that *p* and that if *p* then *q* are the reasons for which you believe that *q*, then as far as (WSO) is concerned, things have gone as well as they can, and you're justified in believing that *q*.

But your acceptance of (WSO) need not ever be one of the reasons for which you believe anything. You can be reasonable in virtue of fulfilling this requirement without having to represent it. You don't need a linking mental state. One thing that matters from the point of view of rationality is how good your reasons really are, not just how good you think they are. Another thing that matters is that you avoid the combinations that really are bad, not just the ones that look bad to you. (WSO) is a rational requirement. You don't have to like it; you don't have to accept it; but you do have to comply.

And when things go as well as they can, you get more than mere compliance, if that just means acting in accord with the rule. If your belief that arithmetic is incomplete is based on your belief that the sky is blue, you believe the right thing. But you don't believe it for the right reasons. If you believe that *p*, believe that if *p* then *q*, and also believe that *q*, we can suppose that in believing that *q*, you believe

the right thing. But we don't know yet whether you believe it for the right reasons. But if your justified, first-order beliefs are the reasons for which you believe that q, then you can believe the right thing for the right reasons without having to represent the rule or requirement.

Complaining about no-frills reliabilism, and even no-frills hyper-reliabilism is very tempting. But it's also risky business. You've got this reliable process. So your beliefs are either likely or guaranteed to be true. What more do you want? Maybe what you want is what your inner tortoise wants, not just the existence of a rational connection between premises and conclusion, but the knowledge that the transition is reliable. If that's what you want, you just can't have it. But if you're satisfied with the existence of the rational connection plus the ability to believe the right things for the right reasons, then it no longer looks as though the only options are the tortoise's impossible conditions and mere hyper-reliability.

2.2 Some practical cases

You get the same sort of thing in the practical case. Suppose you have a standing intention to take an umbrella if it's raining. The question of whether you ought to take an umbrella is connected to the question of whether it's raining. In this case, you figure out what to do by figuring out what's true. You answer a question about one thing by thinking about something else. And the questions are connected by way of your conditional intention.

But other cases are messier. You haven't made up your mind about whether it's raining; whether to take an umbrella; or even whether you ought to take an umbrella if it's raining. Many different combinations of answers to these questions are just fine. But some combinations are no good from the point of view of rationality, like believing it's raining; forming the conditional intention; and deciding not to take your umbrella.

Why are these three questions connected? It's not by way of an antecedent acceptance of the principle of instrumental reason whatever exactly that is. And it's not by way of a second-order intention to form the first-order intention if you have the belief. If the first-order intention to ϕ if p plus the belief that p can't get you to ϕ, then the second-order intention to intend to ϕ if you believe that p plus the belief that you believe that p won't be any better off. These questions are connected but not by way of a linking mental state. The reasons themselves get you to avoid the bad combinations, at least when you're being reasonable.

One interesting thing happens when we turn to the practical case. Normative questions sneak in all by themselves. Did you hear it? It's about whether you ought to take your umbrella. Let's start with a two-person case. Suppose that for whatever reason, the question of what you're doing for dinner is not independent of the question of what your mate is doing for dinner. When it comes to speech acts, the following distinction is huge: there's starting the conversation about dinner, and there's inviting them to decide. Certain sentences are typically just as good for one as they are for the other:

Factual: What are we doing for dinner?

Mental: What do you want for dinner?

and

Normative: What should we do for dinner?

In most ordinary circumstances, these sentences are pragmatically indistinguishable. Whichever sentence you pick, you need to make clear what you're doing. Any could be misinterpreted as doing the other thing. And when things go well, any could be used successfully to do either.

But it's not just that you could start the conversation with any of these sentences. It's that in the normal case, people move seamlessly between talking about "oughta," "wanna," and "gonna." And it doesn't seem to them as though they are in any way changing the subject. It's as if there's only one question here and three ways of putting it. And it is as if there's only one question here, and that's important. But it's only *as if*.

It would be very, very bad to move from typical pragmatic indistinguishability to the conclusion that these questions all come to the same thing or that there's a single fact that determines the correct answer to them all. Literally identifying the factual question with the normative question commits you to this:

$$(\text{The Crazy Idea})\Box\,(\phi)\,(\text{we ought to } \phi \text{ iff we're going to})$$

No one's that good.

Let's turn to a one-person case. Now you have to make up your mind on your own what to do for dinner. You can ask yourself the following three questions:

Factual: What am I going to do for dinner?

Mental: What do I want for dinner?

and

Normative: What should I do for dinner?

It's as if there's only one question and three ways of putting it: the practical or deliberative question of what to do for dinner. But if we literally identified the questions, we'd be back to The Crazy Idea in its first-person form. The questions are different, but you have to answer them together because they're connected. And you answer these questions by thinking about food.

There are three important things about this case.

(1) The sentences are typically pragmatically indistinguishable because the questions (the things asked by use of the interrogatives on a particular occasion) are or ought to be answered together.

You move seamlessly from one sentence to another in the ordinary case because you're trying to answer all three questions together. So the connection between the normative and the practical or deliberative question (whichever that turns out to be) is not just a surface feature of ordinary language. It's a feature of practical thinking.

(2) The questions don't have to be connected by way of a linking mental state.

You answer the question of what to do for dinner by forming an intention. But in order to connect this question with the normative question, you don't need a second-order intention to intend to ϕ when you think that you should. And you don't need the antecedent acceptance of the anti-akratic principle in order to engage in practical reasoning.

(3) The questions are not connected by way of a necessary connection between the facts.

So of course it's not true that necessarily, if you believe you ought to ϕ then you intend to ϕ. The question of whether you ought to ϕ is not independent of the question of whether to ϕ. But the connection between the questions is normative and epistemic, not necessary. You ought to be able to answer these questions together. And you will answer these questions together when you're being reasonable. But nobody's reasonable all of the time. And if you can't answer these questions together, there's something wrong with you in the rationality department.

So while certain combinations of answers to these questions are just fine, other combinations are ruled out by the requirements of rationality.

I ought to ϕ, but I'm not going to.

That's pretty much what it is to be akratic. It's to think that you have most good reason to do one thing while intending or doing something else. This is neither impossible nor uncommon. I do it all the time. But it involves you in some kind of irrationality.

Again, we can think of the relevant rational requirement in terms of a wide-scope "ought." Where ϕ-ing is something you're in a position to do for a reason, if the questions of whether you ought to ϕ and of whether to ϕ come up, then your answers to those questions ought to be like this: You believe you ought to ϕ only if you ϕ.

$$O[BO\phi \rightarrow \text{you } \phi]$$

In the case of (WSO), there is a necessary connection between the facts that p and that if p then q on the one hand and the fact that q on the other. And we just can't help but think that the necessary connection between the facts must have something to do with the normative connection between the beliefs. But in the case of the anti-akratic principle, there's no necessary connection between believing that you ought to ϕ and ϕ-ing. So what is wrong with being akratic?

In this case, the questions are connected not because of a necessary connection between the facts, but because of a connection between the reasons. Any reason to think you ought to ϕ is itself a reason to ϕ. And a good reason not to ϕ is a reason to think you shouldn't. So if your answer to the question of whether to ϕ comes apart from your answer to the question of whether you should, then you're being irrational somewhere. Either you're believing or acting for bad reasons.

If the "O"s in the principle express the requirements of rationality, then when you're akratic, you see yourself as irrational. What you do conflicts with your own

sense of the reasons for doing it. But when you are being reasonable, and you do answer these questions together, the reasons themselves can get you to avoid the bad combinations. If you believe you ought to ϕ, you have two options. You can either revise the belief or ϕ. If all the evidence suggests that you ought to ϕ, then you shouldn't revise the belief, and the only permissible option left is to just do it. If your belief that you ought to ϕ is the reason for which you ϕ, then you ϕ for the right reason. And if the things that get you to believe you ought to ϕ also get you to ϕ, then you ϕ for the right reason. But your acceptance of the anti-akratic principle need not ever be your reason for anything.

2.3 A theoretical case

Let's turn to a one-person theoretical case. You might ask yourself the following questions:

(F) Is *p* true?
(M) Do I believe that *p*?
(J) Am I justified in believing that *p*?
(N) Should I believe that *p*?
(K) Do I know that *p*?

You answer the questions (M), (J), (N), and (K) by thinking about the world, i.e., by trying to answer (F). When it comes to answering, (F) seems like the fundamental question. But it would be just crazy to identify the facts.

Sometimes it's okay, and even more than just okay, to answer questions about one thing by thinking about something else. And this is okay when the questions are connected. The questions don't have to be connected by way of a linking mental state, and they don't have to be connected by way of a necessary connection between the facts. The questions can be connected by way of the reasons. It's okay to answer questions about one thing by thinking about something else when thinking about that something else gives you reasons to answer the questions one way or another.

It's not that you always ought to give the same answers to all of these questions. Is there an even number of trees in Hyde Park? I don't know. Do I believe there's an even number? Absolutely not. Like Modus Ponens and akrasia, only certain combinations are ruled out.

So which combinations are ruled out in our theoretical case? There are a lot, but here's a place to start.

(O) It's raining, but I don't believe that.

and

(C) I believe it's raining, but it's not.

It's not just that Moore-paradoxical beliefs involve giving different answers to (F) and (M). It's that there's a rational conflict between these two answers (Moore 1952). Moore-paradoxical beliefs are irrational or incoherent. That's supposed to be

obvious. The hard question is why they're incoherent given that the proposition believed is contingent.

If there's no necessary connection between the facts, what is wrong with believing Moore-paradoxical things? Here's one way of thinking about it (Gibbons (2013). In both cases, you see yourself as irrational. Suppose you believe (O). Either you have sufficient evidence for the first conjunct, or you don't. If you do have reason to believe that it's raining, that is reason to believe that the second conjunct shouldn't be true. If it's true when it shouldn't be, then you're irrational. If, on the other hand, you don't have evidence for the first conjunct but believe it anyway, then you're irrational.

Now suppose you believe (C). Either you have sufficient evidence for the second conjunct, or you don't. If you do have reason to believe that it's not raining, that is reason to believe that the first conjunct shouldn't be true. If it's true when it shouldn't be, then you're irrational. If, on the other hand, you don't have evidence for the second conjunct but believe it anyway, then you're irrational.

The idea is not that you could never be justified in believing both conjuncts. If your therapist convinces you that deep down, you believe your mom is out to get you, you may know full well that she's not out to get you. So you could know something like (C). But you need your therapist to arrive at the second-order belief because your repressed first-order belief about your mom is not responding to reasons, and it's not behaving in the way that good reasons should. And that's why you can't know about it in the ordinary, first-person way.

The first-order belief that p is itself a reason to believe that you believe that p. It's the very best reason to believe that you believe that p. If you form the second-order belief in the ordinary, normal way, you get not only knowledge but also a special status. The second-order belief is justified on the basis of the fact that it's about. There's nothing wrong with taking other routes. But if the ordinary route up is not available to you, there's something wrong with you in the rationality department.

The real problem with the observation model of self-knowledge is that the model leaves out the demand for rational integration of the first and second orders (Shoemaker 1996; Burge 2000; Moran 2001). One of the most distinctive things about self-knowledge is that what's thought and thought about depend on the same point of view. The normative significance of this is that your first- and second-order beliefs should be responsive to the same set of reasons. The requirements of integration can be thought of in terms of wide-scope "oughts." Here are some things to consider, one for (O) and one for (C).

(WO) $\mathrm{O}\,[\mathrm{B}p \rightarrow \mathrm{BB}p]$
(WC) $\mathrm{O}\,[\mathrm{BB}p \rightarrow \mathrm{B}p]$

The conjunction of these two looks a little harsh. (WO) seems to be saying that you should leave no stone unturned. Form the second-order belief whenever you have the first-order belief. And (WC) seems to be saying that you should never get it wrong when it comes to beliefs about your own mind.

The easiest way to think about all of the wide-scope "oughts" that we're considering is that they don't really require you to have a view. What this pair says, in effect, is that if you do answer the question of whether p is true and you also

answer the question of whether you believe it, you should answer these questions together, and your answers should be related like this: you believe that p if and only if you believe that you believe that p. This is an expression of the idea that the ordinary route up ought to be available to you, not that it necessarily will be.

> Here's another combination that Moore thought was Moore-paradoxical (1962).
>
> It's raining, but I don't know that.

At least on the surface of ordinary language, it's perfectly fine to move seamlessly between talking about knowledge and truth. I tell you that the store is closing early today, and you immediately ask me how I know. To ask where the keys are, you have to pick a sentence. You might ask, "Where are the keys?" Or you might ask, "Do you know where the keys are?" Your choice of sentence is based on considerations of style. It's as if there's only one question here and two ways of putting it. But it's only *as if*. Whether it's a one-person or two-person case, the question of whether it's true is not independent of the question of whether you're in a position to know.

These questions are not connected by a linking mental state or a necessary connection between the facts. They're connected because the reasons are connected. Any reason to believe that you're not in a position to know that p is itself a reason to revise the belief that p (Gibbons 2013). This is how undermining defeaters work. If you find out that half of the things in your neighborhood that look like barns are really barn facades, you're no longer justified in believing that the thing in front of you is a barn.

The thing about the facades doesn't show that your belief was false or unjustified. But it does show that there was something so seriously wrong with it from the epistemic point of view that you need to revise it. And what it shows is that you weren't in a position to know. If reason to believe you don't know whether or not p is reason to withhold, then it's not just an accident that we use the sentence we do to express the state of seriously withholding judgment: I don't know.

There's no problem with my believing that it's raining but you don't know that. But there is a problem with your believing that it's raining but you don't know that. And the problem is fairly straightforward. Any reason for you to believe the second conjunct is reason for you not to believe the first. Your first-order and second-order beliefs ought to be responsive to the same set of reasons. And here's a way of thinking about one of the requirements on how these things ought to be connected.

$$\text{(WK)} \quad \mathrm{O}\,[\mathrm{K}p \rightarrow \mathrm{BK}p]$$

If this were a narrow-scope "ought," we might get some kind of regress. But in any case, it would be another version of the idea that you should leave no stone unturned. But if we read (WK) on the model of (WO), it just tells you how your answers to these questions ought to be organized at the end of the day, if you're going to answer them.

In the first instance, wide-scope "oughts" rule out the bad combinations. Many wide-scope "oughts" rule out the worst-case scenario. Some of them also rule out the second-worst-case scenario. But they also bring with them an implicit conception of the best-case scenario. One idea behind (WSO) is the idea that the beliefs that p and that if p then q are reasons to believe that q. And if they're good reasons, these reasons can be good enough on their own. You don't need to be justified in believing the tortoise's first premise or in accepting either Modus Ponens or (WSO). If your justified first-order beliefs are the reasons for which you believe that q, then from the point of view of rationality, this is as good as it needs to get.

One idea behind the anti-akratic principle is the idea that the belief that you should is a reason to do it. And if this reason is good enough, it's good enough on its own. You don't also need to accept the principles, or the rules, or what have you. So you could intend, believe, and do the right things for the right reasons without having to meet the tortoise's impossible conditions.

One idea behind (WK) is the idea that the knowledge that p is a good reason to believe that you know that p. And in fact, it's the very best reason to believe that you know. When the belief that you know is justified on the basis of the fact it's about, you get the kind of self-knowledge that comes with a special status. And in the ordinary, normal case, when you're being reasonable, you have available to you a distinctive route up from first-order to second-order knowledge. If this route is not available to you, there may well be something wrong with you in the rationality department.

When we think about self-knowledge in terms of counting kinds of error possibilities, it looks like there's a huge difference between knowing what we believe and knowing that we know. But there's another picture of how self-knowledge works. Introspective knowledge is based on what it's about. And the demand for rational integration of the orders explains why the first-order state is a reason for the second-order state. And the picture works just as well for knowledge and belief.

The picture explains not only why the route confers the status and all that. It also explains why the first-order question about the world is the fundamental question, at least when it comes to answering them. When things go well, what justifies your belief that you believe that p is your belief that p. And what justifies your belief that you know that p is your knowledge that p. And so on. So in these sorts of cases, in order to be justified in believing the second-order thing, you need to be in the first-order mental state. And being in the first-order mental state just is thinking about the world.

The key to knowing that you know is the key to good cooking. Start with good ingredients, and don't mess anything up. If you know that p, you're starting with good ingredients. You've got truth, justification, safety, and all that. It's not that necessarily, in every possible case if you can't figure out from here whether or not you know, it's automatically your fault. It's that there's a strong presumption in the ordinary case that if you can't get there from here, you've probably messed something up. When it comes to cooking, not messing anything up requires a certain amount of skill. But knowing that you know is so much easier. All you have to do is

answer the question about the mind by thinking about the world. And that's what we do every day.

Acknowledgements I would like to thank Casey Doyle, John Hawthorne, Thomas Hofweber, Ram Neta, participants in the Aims and Norms workshop at the University of Oslo, audiences at UNC Chapel Hill and Oxford, and an anonymous referee for *Philosophical Studies* for helpful comments and suggestions.

References

Alston, W. (1971). Varieties of privileged access. *American Philosophical Quarterly, 8*(3), 223–241.
Anscombe, G. E. M. (1957). *Intention*. Oxford: Blackwell.
BonJour, L. (1985). *The structure of empirical knowledge*. Cambridge, MA: Harvard University Press.
Broome, J. (1999). Normative requirements. *Ratio, 12,* 398–419.
Burge, T. (2000). Reason and the first person. In C. Wright, B. Smith, & C. Macdonald (Eds.), *Knowing our own minds*. Oxford: Oxford University Press.
Byrne, A. (2005). Introspection. *Philosophical Topics, 33,* 79–104.
Byrne, A. (2011). Knowing what I want. In J. Liu & J. Perry (Eds.), *Consciousness and the self: New essays*. Cambridge: Cambridge University Press.
Carroll, L. (1895). What the tortoise said to achilles. *Mind, 4,* 278–280.
Davidson, D. (1963). Actions, reasons, and causes. In *Essays on actions and events*. Oxford: Clarendon.
Davidson, D. (1986). A coherence theory of truth and knowledge. In E. LePore (Ed.), *Truth and interpretation: Perspectives on the philosophy of Donald Davidson*. Oxford: Blackwell.
Edgley, R. (1969). *Reason in theory and practice*. London: Hutchinson.
Evans, G. (1982). *The varieties of reference*. Oxford: Clarendon.
Gibbons, J. (1996). Externalism and knowledge of content. *Philosophical Review, 105,* 287–310.
Gibbons, J. (2010). Seeing what you're doing. In T. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (Vol. 3, pp. 63–85). Oxford: Oxford University Press.
Gibbons, J. (2013). *The norm of belief*. Oxford: Oxford University Press.
Humberstone, L. (1992). Direction of fit. *Mind, 101,* 59–83.
Leite, A. (2008). Believing one's reasons are good. *Synthese, 161,* 419–441.
Moore, G. E. (1952). Reply to my critics. In P. Schilpp (Ed.), *The philosophy of G. E. Moore*. New York: Tudor Publishing Co.
Moore, G. E. (1962). *Commonplace book: 1919–1953*. London: Allen & Unwin.
Moran, R. (2001). *Authority and estrangement*. Princeton: Princeton University Press.
Neta, R. (2011). The nature and reach of privileged access. In A. Hatzimoysis (Ed.), *Self-knowledge*. Oxford: Oxford University Press.
Peacocke, C. (2000). Conscious attitudes, attention, and self-knowledge. In C. Wright, B. Smith, & C. Macdonald (Eds.), *Knowing our own minds*. Oxford: Oxford University Press.
Pritchard, D. (2005). *Epistemic luck*. Oxford: Oxford University Press.
Shoemaker, S. (1996). *The First-Person Perspective and Other Essays*. New York: Cambridge University Press.
Sosa, E. (1980). The raft and the pyramid: Coherence versus foundations in the theory of knowledge. *Midwest Studies in Philosophy, 5,* 3–25.
Sosa, E. (1999). How to defeat opposition to Moore. *Philosophical Perspectives, 13,* 141–154.
Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.
Wright, C. (1985). Facts and certainty. *Proceedings of the British Academy, 71,* 429–472.