Critical Reasoning and The Inferential Transparency Method

Abstract: Alex Byrne (2005, 2011a, 2011b, 2018) has argued that we can gain self-knowledge of our current mental states through the use of a *transparency method*. A transparency method provides an extrospective rather than introspective route to self-knowledge. For example, one comes to know whether one believes *p* not by thinking about oneself but by considering the world-directed question of whether *p* is true. According to Byrne, this psychological process consists in drawing inferences from world-directed propositions to mind-directed conclusions. In this paper I consider whether his 'Inferential Transparency Method' can provide us with the self-knowledge that many philosophers think we require in order to "critically reason" about our mental states (Burge 1996). I argue that the Inferential Transparency Method cannot provide this. Whether this is a damning objection for Byrne depends on how much stock we should place in our status as critical reasoners. However, I conclude by suggesting a more general worry for Byrne's account.

Keywords: Self-Knowledge; Transparency; Inference; Taking Condition; Rule Following

§I—Introduction

Alex Byrne (2005, 2011a, 2011b, 2018) has meticulously argued that we can gain self-knowledge of our current mental states by means of a *transparency method*. Like many others, Byrne draws his inspiration from Gareth Evans, who once wrote that:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me "Do you think there is going to be a third world war?" I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question "Will there be a third world war?" (1982, p. 225)

Setting aside Evans' probably overly strong claim that we *must* gain self-knowledge along the above lines, his insight (if it is one) is that self-knowledge *can* be an extrospective rather than introspective achievement. Evans' claim, in other words, is that the mind is *transparent to the world*: how things are with one's mental states can be discerned by "seeing through" to what one's mind regards. Byrne's contribution to our understanding of this idea is an account of the psychological process that produces transparent self-knowledge. He argues for a *transparency method* based on inferences from world-directed premises to mind-directed conclusions. I will refer to this as Byrne's *Inferential Transparency Method*, or ITM.

Byrne offers ITM in order to both flesh out Evans' suggestion and explain the seemingly *privileged* and *peculiar* features of much of our self-knowledge. Roughly, self-knowledge is privileged to the extent that it is especially epistemically secure relative to other varieties of empirical knowledge, and it is peculiar to the extent that it is acquired by means available only to the subject herself. It has often been objected that transparency methods can only explain the privilege and peculiarity with which we know a subset of our current mental states (Finkelstein 2003; Bar-On 2004; Gertler 2011b; Borgoni 2018). One popular version of this objection is that not all mental states have worldly contents that one can grasp before self-ascribing one's mental state. Byrne (2018) tackles this objection head-on by developing an impressive array of transparent inference schemas suited to delivering self-knowledge of different kinds of mental states. But even if he is unsuccessful here, we might simply embrace pluralism about the sources of privileged and peculiar self-knowledge (Moran 2001; Boyle 2009; Samoilova 2016; Coliva 2016; Parent 2017; Komorowska-Mach 2019).

In this paper I want to take up a different concern about ITM (though I will briefly revisit the above objection, slightly modified, in §IV). In doing so, I will mostly focus on ITM as an account of how we can know our own propositional attitudes. My goal is to evaluate whether ITM can yield the sort of self-knowledge that some philosophers think we need in order to *critically reason* about our propositional attitudes. Roughly, to critically reason is to engage in higher-order evaluations about the reasonability of one's lower-order propositional attitudes, with an aim to improving one's overall rationality. I will argue that ITM does not enable us to critically reason, at least given Tyler Burge's highly influential conception of critical reasoning (1996, 2013). Whether this is a damning objection for Byrne depends on how much stock we

should place in our status as Burgean critical reasoners.[1] However, I will conclude this paper by suggesting a more general worry for Byrne's account.

Here is the plan for what follows. In §II, following Burge and his recent interpreter Ben Sorgiovanni, I say more about what it is to be a critical reasoner, and I explain why these authors take critical reasoning to require *first-personal* self-knowledge. What makes self-knowledge first-personal is shown to include but not reduce to its privilege and peculiarity. I also briefly touch upon Burge's *epistemic* account of self-knowledge, which is a transcendental account of the warrant our beliefs about our own mental states (hereafter *self-beliefs*) enjoy as critical reasoners. This leaves us with a need for a *psychological* account of first-personal self-knowledge, i.e., an account of the psychological process (if there is one) through which we acquire first-personal self-knowledge. I call this *the supplementation question*, since it is the question of how best to supplement Burge's account of the relationship between critical reasoning and first-personal self-knowledge with a psychological story of how we acquire the latter. In §III I explicate ITM. After considering some objections to ITM I conclude that, prima facie, ITM provides a compelling answer to the supplementation question. However, in §IV I change gears and argue that ITM nevertheless fails to answer it. In §V I conclude by suggesting that, critical reasoning aside, ordinary agents likely do not use ITM to acquire self-knowledge.

§II—Critical Reasoning and First-Personal Self-Knowledge

In critical reasoning, agents aim to conform their attitudes to rational norms, and to evaluate the reasoning that produces their attitudes.[2] It is reasoning "guided by an appreciation, use, and assessment of reasons and reasoning as such" (1996, p. 98). We critically reason, Burge says, in

[1] For skepticism, see Owens (2000, 2011) and Cassam (2015).
[2] Because my main aim is not to defend Burge, my exegesis of his account will sometimes be highly compressed. For more detailed treatments of many features of Burge's account, see Sorgiovanni (2018).

the course of "…giving a proof, in thinking through a plan, in constructing a theory, in engaging in a debate" and so on (1996, p. 99). These are ubiquitous cognitive activities, at least among adult human agents. No wonder, then, that Burge takes a capacity for critical reasoning as partly constitutive of fully developed human personhood (Burge 2013).

Here is a toy example. Imagine that Ev is concerned about recent trends in her country's political discourse. In search of an explanation, she works her way into the following thought process. First, she notices that her belief P (*that her country is safe from fascist takeover*) squares poorly with her recently acquired and evidentially well supported belief Q (*that fascist rhetoric is becoming more popular in her country's political discourse*) and her further, equally evidentially supported belief R (*that fascist movements are quickly gaining traction in nearby countries*). She acknowledges that she is rationally required to believe in accordance with her evidence, and eventually concludes that she ought not to believe that her country is safe from fascist takeover. To the extent that this self-belief is itself warranted, it exerts genuine rational pressure on her to disbelieve P. In Burge's words, "justifiably finding one's reasons invalid or one's thoughts unjustified, is normally *in itself* a paradigmatic reason…to alter them" (1996, p. 110). *Ceteris paribus*,[3] Ev will excise her P-belief from her psychology. In this way, critical reasoning enhances her first-order rationality.

Burge's most important contribution to our understanding of critical reasoning is an argument for what sort of self-knowledge one must have if critical reasoning is to be reasonable in the first place. His thesis is that the perspective *from which* we critically reason and the perspective *on which* we critically reason must enjoy an "immediate rationally necessary connection" (1996, pp. 109-110). Roughly, this means having critical self-beliefs that are true

[3] Borgoni & Luthra (2017) defend the possibility of epistemic akrasia.

whenever warranted, such that they count as self-knowledge whenever they are warranted. Thus, when Ev critically reasons and concludes that she ought not to believe that P, her self-belief is warranted only if she really ought not to believe that P, all things considered.[4] I will elaborate on this picture in two broad steps. First, I will clarify this immediate rationally necessary connection between higher- and lower-order perspectives and explain why certain psychological accounts of self-knowledge cannot produce such a connection. After this, I will clarify why Burge thinks that critical reasoning must be conducted only when such a connection obtains.

To clarify the nature of this connection, Burge points the reader to cases where one has a warranted, critical higher-order belief, but where this belief is *not* related in an immediate rationally necessary way to its lower-order object.[5] Paradigm cases are critical beliefs about *other* people's propositional attitudes. For example, I might believe that Pete believes that the stovetop is hot (I infer this from watching him hesitantly approach it). Next, I judge that he ought not to believe that, because the stove broke down last night. It can transpire, however, that "[I may be] in error about what [Pete's] beliefs are, or [Pete's] perspective may have different associated reasons or background information from mine" (Burge 1996, p. 109); perhaps, for example, Pete is only hesitant because he believes there is a spider on the stovetop. This is a basic fact about my epistemic relationship to Pete: this dissociation between my beliefs and Pete's, despite the reasonability of my beliefs, entails that there can always be a gap between my warranted beliefs about what Pete ought to believe and the facts about what Pete ought to believe. In other words, what I believe about what Pete ought to believe lacks immediate rationally necessary consequences for what Pete ought to believe.

[4] I won't continually reissue this 'all things considered' qualifier going forward.
[5] I will focus on critical reasoning about first-order mental states, as opposed to second-order or higher mental states.

But now note that, if I ascribe a mental state to *myself* on the same basis, I will stand to my mind in a highly similar way as I stand to Pete's. By observing my behaviour, judging that I believe P, and inferring that I lack good reason to believe that P, my self-belief can be warranted yet false in the same way.[6] Hence, my self-belief does not have immediate rationally necessary consequences for whether I ought to believe P. Burge's suggestion is that similar lessons apply to many other ways of forming critical self-beliefs: if my beliefs about my attitudes are based on Pete's testimony, or on inductive inferences about the likelihood of my believing various things given, e.g., statistical generalizations about men my age, the same opportunities for dissociation between my warranted self-beliefs and their lower-level objects will be present. These methods can be characterized as *third-personal* because they are methods that anybody else might use to acquire knowledge of another's mind. Third-personal methods cannot yield self-beliefs that are related to their objects in an immediate rationally necessary way. In order to critically reason, I must be able to take up a self-perspective that is essentially epistemically different from the perspective of another agent.

Importantly, Burge's critical target encompasses more than third-personal methods for acquiring self-knowledge. Thus, at one point he considers the possibility that we acquire self-knowledge through "*sensed* inner goings-on" (1996, p. 104). Take, for example, the view that self-knowledge is delivered by the operations by operations of a scanning mechanism in the brain (Armstrong 1968; Lycan 1996). Such a scanner, if real, could provide self-knowledge that is *peculiar* so long as it cannot be used to scan another's mental states. Such a scanner, we might also stipulate, could be especially reliable and so account for the *privileged* status of our self-

[6] Granted, it may be true that we can know ourselves better than others because we have more behavioural evidence about ourselves than others do (Ryle 1949). But self-beliefs derived from such evidence can be warranted yet false in just the way that my beliefs about Pete, similarly derived, can be.

knowledge. Nevertheless, Burge rejects inner sense accounts as answers to supplementation question. This is because it is possible for an inner sense faculty to operate reliably but faultily, and so to deliver warranted false positives. Put differently, these accounts leave open the possibility of "*brute local error*" (Bar-On 2004, p. 98): errors that are not due to any psychological or epistemic failing on the agent's part. In the case at hand, one's scanner may (*qua scanner*) simply misread its object, even if one is cognitively well-functioning, thereby delivering a warranted yet false self-belief. But this means that there is no immediate rationally necessary connection between the self-beliefs produced by one's inner scanner and one's first-order perspective, even though one's inner scanner can also produce warranted and true self-beliefs (Sorgiovanni 2018, p. 5). This is because, in these cases, one only *happens* to have a warranted yet true self-belief—it could have been otherwise, and so the connection between warrant and truth is not necessary. Let us say, then, that first-personal self-knowledge is self-knowledge that is not only privileged and peculiar, but that is also connected to the mental states known in a rationally immediate rationally necessary way. So understood, Burge's criticism is that "inner sense" cannot produce first-personal self-knowledge, even though it can produce privileged and peculiar self-knowledge.

Of course, all of this is interesting only if Burge is right that critical reasoning really does require an immediate rationally necessary connection between one's self-beliefs and their objects. Here, however, one might be skeptical. After all, it can seem like we critically reason in situations where no such connection obtains. For example, I might find myself in some circumstance and come to believe correctly, through a warranted inference about what I typically desire in similar circumstances, that I desire to φ. I might then believe that I ought not to desire this. If I subsequently forfeit my desire on the basis of this belief, which certainly seems

possible, my rationality will be enhanced. But because inferences from my observed behaviours to their mental causes can be warranted yet false, there is *not* an immediate rationally necessary connection between my first- and second-order perspectives in this case. The skeptic now asks: why is such reasoning not good enough to count as bonafide critical reasoning?

In reply, I think we should follow Sorgiovanni's (2018) interpretation of Burge. According to his interpretation, it is *a constitutive norm of critical reasoning itself* that there must be a necessary, immediate rational connection between one's critical and lower-order perspectives. As evidence for this interpretation, Sorgiovanni cites Burge's claim that "it is constitutive of critical reasoning that if the reasons or assumptions being reviewed are justifiably found wanting by the reviewer, it *rationally follows immediately* that there is a *prima facie* reason for changing or supplementing them" (1996, p. 109). His reading of Burge's use of 'constitutive' is as referring to this constitutive norm. According to this norm, having a warranted self-belief that you ought to believe P requires you to believe P. Now, the reason why there should be such a norm is precisely that it is only when such a norm is fulfilled that critical reasoning inherently adds to the reasonability of one's lower-order mental life.

To see this, we need only reflect on the fact that critical reasoning *unbound* by such a norm would allow for situations in which a self-belief was warranted yet false, whereupon one would be required to change one's mind in a way that would *not* enhance one's lower-order rationality. It is because such possibilities are antithetical to the inherent reasonability of critical reasoning that the whole enterprise is constitutively structured by a norm that requires us to avoid them. If critical reasoning is to add to the reasonability of our reasoning and attitudes, it must be guided by such a norm, even if only in the self-directed case (Sorgiovanni 2018, p. 10).[7] This is why,

[7] Plausibly, no such norm holds for critically reasoning about other minds (Sorgiovanni 2018, p. 11).

even if critical reasoning that violates this norm is possible, it remains the case that such reasoning violates this norm and, hence, does not essentially contribute to my improved lower-order rationality.[8] Indeed, unless one meets this norm in critical reasoning, one's critical self-perspective will not be essentially different from that of another agent's, for we have already seen that another agent's beliefs about what attitudes one ought to have do *not* necessarily bear on what attitudes one ought to have (Sorgiovanni 2018, p. 12).[9]

The final piece of Burge's account is an *epistemic* account of self-knowledge, that is, an account of the entitlement we have to our critical self-beliefs. Here is Burge:

> if one lacked entitlement to judgments about one's attitudes, there could be no norms of reason governing how one ought check, weigh, overturn, confirm reasons or reasoning…If reflection provided no reason-endorsed judgments about the attitudes, the rational connection between the attitudes reflected upon and the reflection would be broken. So reasons could not apply to how the attitudes should be changed, suspended, or confirmed on the basis of reasoning depending on such reflection. But critical reasoning just is reasoning in which norms of reason apply to how attitudes should be affected partly on the basis of reasoning that derives from judgments about one's attitudes. So one must have an epistemic entitlement to one's judgments about one's attitudes. (1996, pp. 101-102)

This argument is transcendental. Because (1) we are critical reasoners, and (2) critical reasoning is itself reasonable, it follows that (3) the self-beliefs that constitute our critical self-perspective must be warranted. We therefore have an "entitlement" to our critical self-beliefs. An entitlement

[8] Gertler (2016) argues that first-personal self-knowledge isn't required for us to actually change our minds in accordance with our critical self-beliefs. But Sorgiovanni (2018, p. 7) points out that such accounts will still violate a constitutive norm of critical reasoning.

[9] An anonymous reviewer imagines a case in which one has excellent but misleading evidence that there is poison in one's cup. The reviewer asks why this evidence is not a reason "in and of itself" for one to not drink from the cup. The case is supposed to be unsettling because it involves warranted yet false evidential beliefs that still seem to serve as reasons in and of themselves to do something, contra what Burge allows for critical self-beliefs. My response is that, if it is true that one's misleading evidence really can serve as a reason in and of itself (in the relevant sense) to not drink from the glass, this is because one's evidential beliefs are part of the same (first-order) perspective as one's other beliefs about the glass and about what one should do with it. Contrariwise, it is because our first- and second-order perspectives have fundamentally different objects that, for Burge, a stringent normative relation is needed to ensure that they are rationally responsive to one another in the right way.

is a species of epistemic warrant that "need not be part of the repertoire of the individual that has the entitlement" (Ibid., p. 94), meaning that it need not be articulable by its possessor.

This is all I will say to explicate Burge's views of critical reasoning and its requisite first-personal self-knowledge. Going forward I will refer to this package of views as *Burgean Agentialism*, since it is a transcendental account about the relationship between a particular form of epistemic agency and self-knowledge. Having identified certain psychological accounts of self-knowledge as unable to produce first-personal self-knowledge (that is, answer the supplementation question for Burgean Agentialists), I will now explore the possibility that Byrne's Inferential Transparency Method (ITM) can fare better.

§III—The Inferential Transparency Method

Byrne develops ITM "using the apparatus of epistemic rules" (2018, p. 102). That is, he takes the use of different epistemic rules to provide routes to self-knowledge for different kinds of mental states. For now, I will mostly focus on what strikes me as ITM's best epistemic rule, namely:

(BEL): if P, believe that you believe that P. (2018, p. 102)

Byrne claims that following (BEL) amounts to performing an inference. Any such inference respects Evans' original insight because one moves from a first-order, world-directed premise to a self-ascription of a belief—a belief that one has whenever one judges that the antecedent of (BEL) obtains.[10]

Now, the fact that (BEL) delivers self-knowledge through an inference might suggest that it cannot deliver first-personal self-knowledge. This is because, as a general matter, reasonable inferences can yield false conclusions. Why, then, should (BEL)-inferences fare any better? A

[10] Byrne sometimes talks of judgement instead of belief. This is fine so long as we understand judgements as (or as entailing the existence of) occurrent beliefs. Peacocke (1998) denies this, though see Zimmerman (2006) for a reply.

first reason, repeatedly stressed by Byrne, is that one can acquire self-knowledge even when one unsuccessfully *attempts* to follow (BEL), which is what happens when P is false and one self-ascribes a belief in its truth anyway. This is because, in these cases, self-ascribing one's false first-order belief does not falsify one's resultant self-belief; instead, one is just self-ascribing one's actual, false first-order belief. This means that (BEL) is "self-verifying" (2018, p. 104): the mere attempt to follow it yields true self-beliefs. One might conclude, therefore, that (BEL) yields privileged self-knowledge because it is a hyper-reliable inference rule. One might also conclude, for the same reason, that it cannot yield warranted yet false self-beliefs, since it yields true self-beliefs even when one merely attempts to follow it.

A second reason for optimism is that, as Byrne also stresses, (BEL) delivers peculiar self-knowledge because it cannot be used to acquire knowledge of other minds. To see this, consider the fact that following a rule like <*(Pete-BEL)*: *if P, believe that Pete believes P*> would not be self-verifying or even reliable, because my first-order beliefs can easily diverge from Pete's. So (BEL) also seems to provide peculiar self-knowledge. Taken together, it can seem that ITM provides privileged and peculiar self-knowledge that cannot yield warranted yet false self-beliefs. Accordingly, it can seem to produce first-personal self-knowledge, and to thereby answer the supplementation question.

Matthew Boyle has argued, however, that (BEL)-inferences are actually unwarranted, since any such inferences would be "mad" (2011, pp. 230-231). This is because (BEL)-inferences are neither deductively valid (P does not entail that I believe it) nor inductively strong (there are infinitely many first-order beliefs that I lack).[11] In reply, Byrne simply reiterates that the self-verifying nature of (BEL)-inferences makes them warranted despite their lacking other canonical

[11] Sorgiovanni (2018, en. 6) follows Boyle here and so denies that ITM answers the supplementation question.

good-making properties of inference. In other words, it is because self-beliefs derived from following (BEL) are maximally "safe"—in the sense that they could not have easily been false—that they count as self-knowledge.[12] Alternatively, it might be possible to argue (in our dialectical context at least) that, so long as one retains a Burgean entitlement to one's self-beliefs (see §III), it is irrelevant whether Byrne's *psychological* account of self-knowledge answers *epistemic* questions about nature and source of the warrant we have for our self-beliefs.

A related charge, also raised by Boyle, is that an agent who follows (BEL) cannot explain why, from her own perspective, her self-belief is warranted. The charge is related because, as Boyle reminds us, P is not good evidence for the fact that one believes P, nor does P deductively entail that one believes that P, and so the agent herself cannot appeal to any such evidence or entailment to epistemically ground her self-belief. In reply, however, Byrne notes that this is exactly what we should expect. For it is generally agreed by epistemologists that self-knowledge is not based on evidence, at least from the first-person perspective.[13] Indeed, as we saw in §II, Burge embraces a similar point when he says that, despite one's being entitled to one's self-beliefs, this entitlement need not be articulable to its possessor: "[m]ost of us have no justifying argument or evidence backing the relevant self-judgments" (1996, p. 94).

For these reasons, Boyle's arguments against (BEL) appear ineffective. This means that we have not yet seen any reason to denounce (BEL)—and perhaps ITM more generally—as an answer to the supplementation question. One final virtue of ITM is also worth noting, which is that supplementing Burgean Agentialism by way of ITM allows for a psychologically economical conception of the overall process of critical reasoning. This is because the process

[12] Byrne says that "[s]afety is a plausible necessary condition for knowledge." Moreover, "provided it is emphasized that the relevant sense of 'could not have easily been false' cannot be elucidated in knowledge free terms, there is no obvious reason to suppose that it is not also sufficient" (2018, p. 110).
[13] See, e.g., (Davidson 1984; Wright 2001; Bar-On 2004).

begins with one distinctive kind of (bottom-up, world-to-mind) reasoning that provides us with first-personal self-knowledge, and culminates in another distinctive kind of (top-down, mind-to-world) reasoning that enhances the rationality of the mind thus known.[14]

§IV—ITM and the Supplementation Question

In this section I will argue that, despite the appearances, ITM still cannot answer the supplementation question. To begin, let me first address some concerns about *just how much* supplementation ITM might provide. For reasons that I will not belabor here, Byrne believes that we must have a single account of privileged and peculiar self-knowledge.[15] Now, what if we also suppose that we must have a single account of *first-personal* self-knowledge? The trouble at this point is that there are no transparent epistemic rules besides (BEL) that stand out as good candidates for delivering first-personal self-knowledge. To take an example, I may look to the world and discern that φ-ing would be desirable, thereafter self-ascribing a desire to φ.[16] Alas, while "[o]ne's desires *tend to* line up with one's knowledge of the desirability of the options", Byrne himself admits that they *merely* tend to so align (2018, p. 161, emphasis mine). In other words, warranted yet false self-ascriptions seem possible here. This is relevant because, despite having focused primarily on (BEL) as a route to first-personal self-knowledge of belief, we can also critically reason about many of our desires and other propositional attitudes.

Perhaps a similar concern is more pressing, which is that ITM does not even constitute a complete psychological account of first-personal self-knowledge of belief. I have in mind the concern that ITM cannot explain how we can know those of our beliefs that we take, from our

[14] It may be, therefore, that ITM provides a more economical answer to the supplementation question than Sorgiovanni's preferred answer (2018, section 4), which appeals to a seemingly *sui generis* reflective capacity.
[15] Briefly, his concern is that a "uniform" explanation is the only way to explain why we are not subject to "dissociations" in which we know some states in a privileged and peculiar way but not others (2018, pp. 157-158).
[16] Byrne's transparent inference schema for desire: "DES: If ϕing is a desirable option, believe that you want to ϕ".

own perspective, to be unreasonable.[17] This is because ITM yields self-knowledge of beliefs that are grounded in our own judgements—since (BEL)-inferences begin from our first-order judgements—and so the question arises of how self-knowledge of a belief that is based on one's own judgement can strike one as unreasonable.[18]

In reply, we might concede the point and simply embrace the need for a pluralist treatment of first-personal self-knowledge, even as regards belief. Alternatively, focusing on the case of belief, we might mitigate the objection's force in the following way. To begin, note that many of our unreasonable beliefs seem to lack judgement-sensitivity. For example, I may believe in ghosts on the basis of a childhood trauma, even though I would not now judge on any evidential basis that there are ghosts. So, my belief will strike me as unreasonable by my own lights. However, this belief may be also be recalcitrant to critical reasoning, since I already judge it to be unreasonable and yet it persists. But then it may be useless to critically reason about it, because it is not an attitude that would be sufficiently sensitive to critical reasoning in the first place. Admittedly, this reply leaves us with the puzzle of how we can ever know our sufficiently judgement-sensitive albeit unreasonable (by our lights) beliefs via ITM. I am not sure whether ITM can close this gap, although there is a tradition which avoids the problem by rejecting the claim that we cannot even have privileged or peculiar self-knowledge of such beliefs, such that we cannot have first-personal self-knowledge of them either.[19]

Finally, it is worth noting that ITM can still provide us with a self-perspective from which we can *eventually* come to view many of our beliefs as unreasonable. For instance, I might

[17] Thanks to Brie Gertler for this objection, which occurs in Gertler (2011a, chapter 6) and Leite (2018).
[18] Borgoni also argues that no transparency method captures "cogito-like" cases of self-knowledge like *I am hereby thinking/judging that P*. Cogito-like cases are interesting in that they are made true "*in* and *through*" their being thought (2018, p. 683). Borgoni thinks that "understanding the type of self-knowledge involved in *cogito* judgments is crucial for our understanding of the general phenomenon of self-knowledge" (2018, p. 694), such that one's psychological account of self-knowledge should capture them. If so, ITM is on worse footing than I will argue here.
[19] See Bilgrami (2006, 2012) and Coliva (2016), *pace* Leite (2018). See Borgoni (2015) for an intermediate position.

perform (BEL)-inferences from my judgements P, Q, and R. *Ex hypothesi*, all three of my resultant self-beliefs are based on my own judgements. However, once I have all of these beliefs in view, I may *then* go on to recognize that they form an inconsistent triad and, eventually, critically reason my way to the conclusion that one of them is unreasonable.

This is all I will say about objections that question how much supplementation ITM can provide Burgean Agentialism. With this said, I will now develop a different objection, one that disputes ITM's capacity to supplement Burgean Agentialism *even in the best case*, i.e., the case of judgement-based belief. This objection will take the form of a dilemma: I will be arguing that (1) ITM *can* yield warranted yet false self-beliefs about our judgement-based beliefs, or (2) it presupposes the very self-knowledge it aims to explain.

I begin with an orthodox understanding of inference as a temporally extended mental process. Given this understanding, the following possibility might obtain: in performing a (BEL)-inference, I cease to believe that P in the temporal space between first judging that P and forming my conclusion-belief that I believe P. Call this phenomenon *inferential belief extinction*. When this obtains, my self-belief will turn out to be false upon formation. Now my objection is this: if inferential belief extinction is really possible (though hold out for the second horn of my dilemma if you are skeptical), then (BEL)-inferences *can* produce warranted yet false self-beliefs, since it is plausible that (BEL)'s *reliability* is not compromised by this sort of possibility.

Perhaps surprisingly, given his claim that (BEL) is self-verifying, Byrne himself acknowledges the possibility of inferential belief extinction:

> …since inference is not instantaneous, there is no cast-iron guarantee that one's belief in the premiss will remain by the time one reaches the conclusion, in which case one's belief that one (now) believes that *p* will be false…

But he adds that this is no real threat to (BEL)'s self-verifying status, because:

> …Since the chain of reasoning [via BEL or other transparent epistemic rules] is as short as it gets, this possibility can be ignored. (2011, p. 206; emphasis mine)[20]

However, Byrne's reply constitutes nothing more than a general assurance that maximally short inferences—like (BEL)-inferences—are minimally likely to yield false conclusions due to inferential belief extinction. This is because the worry turns on what Byrne admits to be the temporality of *all* inferences, however short. So, one might agree with Byrne that the possibility of inferential belief extinction can be ignored by (BEL)-users, but the reason for ignoring it can only be that it is *reasonable* for (BEL)-users to ignore it, and *not* that inferential belief extinction is impossible. If so, (BEL)-inferences may yield warranted yet false self-beliefs after all.[21]

Might one produce some principled reason why inferential belief extinction nevertheless undermines our warrant for using (BEL)? Here is how the argument might go. First, we reiterate Burge's claim that (1) the warrant we have for our self-beliefs is grounded in our status as critical reasoners. From here, we argue that (2) inferential belief extinction undermines the conditions necessary for critical reasoning. Given (1)-(2), we can conclude that (3) (BEL)-inferences in which inferential belief extinction occurs are not warranted. This argument seems congenial at first blush. For recall that, according to Burge, the self-beliefs we deploy in critical reasoning must be invulnerable to brute error, and so whenever one's critical self-beliefs are false this must be due to some psychological failing on one's part. Because the subject's psychology is compromised, we can then understand why she might lack a transcendental warrant for critical reasoning. Thus, as long inferential belief extinction is due to some psychological failing, it will be no surprise that its occurrence will undermine our warrant for critical reasoning.

[20] See also Byrne (2018, p. 104).

[21] My objection is not that following (BEL) is not self-verifying because one may erroneously self-ascribe a belief that is *merely a judgement* (cf. Peters 2017, p. 12), since I allow that judgements are a species of belief.

Crucially, for this strategy to work, it must be argued that whatever warrant (BEL) confers on our self-beliefs co-varies with the warrant provided by our Burgean entitlement to these same self-beliefs. For it is only if this condition is met that there can be no cases where my self-belief is warranted by my use of (BEL) despite my lacking a Burgean entitlement to that same self-belief, all while my self-belief is false due to inferential belief extinction. Unfortunately, I do not think this co-variance can be assured.

To begin, let us try to make sense of why a psychological failing must underlie inferential belief extinction. To my mind, the clearest cases involve inferences from *standing beliefs* to self-ascriptions of them. Standing beliefs are diachronically and cross-environmentally stable fixtures of one's belief set. For example, my belief that my name is Ben should remain with me for my whole life barring massive psychological deterioration. It is plausible that the extinction of this belief, whether during an inference or otherwise, would suggest some psychological defect precisely because of what it is to be a standing belief. Plausibly, then, if I infer that I have this belief using (BEL), and my inference turns out false due to inferential belief extinction, this seems like a case where I am not psychologically well-functioning—so much so, it might seem, that I am also surely in too poor of psychological conditions to critically reason.

Now, even if this is right, it so far assures only that I lack a Burgean entitlement to my self-belief. It does not show that I have *no* warrant for my self-belief. The reason is that evaluations of an agent's overall psychological conditions can be made independently from any evaluations of the quality of her reasoning on a particular occasion. In other words, suffering from a psychological defect need not affect the reasonability *of one's reasoning itself* or, in turn, the *product* of one's reasoning, rather than affecting some *other* evaluation of one's broader psychological condition. Take, for demonstrative purposes, the apparent felicity of the following

utterance: "her psychological conditions have deteriorated, but she reasoned perfectly well just there!" Such utterances indicate that we can take people as having epistemically (inferentially) grounded beliefs despite their otherwise poor psychological condition.

To argue that psychological defects necessarily impugn the epistemic warrant one has for one's self-belief would also make it hard to understand the possibility of epistemic akrasia. Thus, imagine that I reason my way to the critical self-belief that I ought not to believe P. Imagine also that I am epistemically akratic with respect to my self-belief: I cannot bring myself to change my P-belief upon self-believing that I ought to. Surely, epistemic akrasia involves some sort of psychological defect. But if this defect necessarily compromises my warrant for my reasoning, then the self-belief I arrive at through my reasoning (and that I am epistemically akratic with respect to) will *not* be warranted. The trouble is that this is self-defeating. For, if my self-belief that I ought not to believe P is unwarranted, then I ought not to change my P-belief after all. But then I am not actually epistemically akratic, since I am not failing to change a belief that I ought to change.[22]

Of course, defective psychological conditions could make a difference to the quality of one's reasoning precisely by constraining the inferences one can draw. Thus, in poor psychological conditions, I might fail to reach a certain conclusion, or I might fail to follow some inferential rule properly. However, these are not the sorts of possibilities involved in (BEL)-inferences that are falsified by inferential belief extinction. In these cases, all that is at issue is that one's premise-belief (upon being inferred from) does not maintain its existence by the time one's conclusion-belief is formed. But is this itself a problem for the reasonability of a (BEL)-inference? I can see no general reason for why it must be: in general, the ontic components of a

[22] One might respond by denying the possibility of epistemic akrasia, though see Borgoni & Luthra (2017) for reasons to reject this move.

process may not endure until the end of that process (think of eating), but we do not automatically conclude that the process is thereby compromised.

So much for the first reason why this 'psychological defect reply' fails. What is interesting is that the argument has not turned on claiming that (BEL)-inferences falsified by inferential belief extinction amount to brute errors. This is because inferential belief extinction may indeed yield false beliefs due to psychological as opposed to brute error. The point is that even in these conditions one's (BEL) inferences may nevertheless yield warranted self-beliefs. It turns out, therefore, that there are potentially more errors than brute ones that critical reasoners must avoid.

The second reason this reply fails is that it takes too narrow a view of the sorts of beliefs that can figure into (BEL)-inferences. This is because the reply only focuses on cases where a (BEL)-inference moves one from a standing belief to a self-ascription of that belief, in order to argue that one's self-belief is unreasonable because of plausible claims about the persistence-conditions for standing beliefs. The problem is that these cases comprise only a range of the cases in which one might follow or attempt to follow (BEL), for one might also (attempt to) infer from *merely occurrent* first-order beliefs to self-beliefs via (BEL). For example, many perceptual beliefs about one's location relative to an object dissipate when one's location relative to the object changes. Thus, forming an occurrent perceptual belief at T_1 may not culminate in a standing disposition to continue affirming it at T_2. Once we recognize that a (BEL)-inference might take us from a merely occurrent belief to a self-belief, the psychological defect reply is on even worse footing. This is because I know of no necessary conditions on how long merely occurrent beliefs must exist for their possessors to count as psychologically well-functioning.[23] So, even if (BEL)-inferences from standing beliefs are unwarranted whenever inferential belief

[23] As Peacocke (2017, p. 366, fn. 28) says: "there is no in-principle limitation on how short-lived genuine beliefs can be."

extinction occurs (though I have disputed even this), (BEL)-inferences from merely occurrent beliefs need not be made by psychologically defective agents if and when inferential belief extinction occurs.

Note that we cannot get around the present concern by simply revising (BEL) as follows:

(STAND-BEL): if P, then, if your P-judgement manifests a standing belief rather than a merely occurrent belief, believe that you believe P.

For, even setting aside the fact, as I have argued, that (BEL)-inferences from standing beliefs can deliver warranted yet false self-beliefs, using (STAND-BEL) *presupposes* that one has self-knowledge, since truly judging the second antecedent means that one is already aware of one's belief, whereas falsely judging the second antecedent means that one's inference might still succumb to (non-psychologically-defective) inferential belief extinction.[24]

With this, we reach the first horn of my dilemma: ITM can yield warranted yet false self-beliefs, and so using it as a precursor to engaging in critical reasoning violates a constitutive norm of critical reasoning.[25]

To avoid this horn, a proponent of ITM might now reply that the preceding arguments depend on a problematic conception of inference and that, if we revise our conception of inference, we will see that inferential belief extinction is not possible in the first place. The idea is that, while there may be no necessary conditions on how long beliefs must exist generally speaking, it may be that beliefs involved in inferences really must persist far enough in time for one to form one's conclusion-belief.

[24] For a similar concern see Barz (2019, p. 8).

[25] An anonymous reviewer has suggested another fix for Byrne, which is to accept a Williamsonian (2000) conception of warrant. On this view, only true beliefs are epistemically warranted, since only true beliefs are entirely epistemically permissible to hold. If Byrne goes this way, this ensures that any false self-belief derived from a (BEL)-inference is unwarranted. However, this move is in tension with Byrne's defense of ITM. This is because it requires denying that (BEL) itself confers warrant on its conclusions in virtue of its self-verifying and safe nature, and so diminishes his painstakingly made case for the claim that (BEL) is a good inference rule.

The only congenial argument I am aware of along these lines comes from recent literature concerning the so-called "Taking Condition" on inference (Boghossian 2014, p. 5). Put roughly, the Taking Condition states that inferentially moving from her premise-belief(s) to a conclusion-belief requires one to actively take one's premise(s) to epistemically support one's conclusion, where "takings" are understood as some sort of higher-order thought, attitude, or mental act.

By quickly describing two recent accounts of what exactly "takings" amount to, we will be able to see how the Taking Condition can potentially prevent the threat of inferential belief extinction. Thus, on one account, an inference occurs when the agent takes her premise(s) to support her conclusion by conferring "inferential force" (Hlobil 2019) on an ordered set of propositions that are all kept in mind at the same time. So, on this view, one's premise-beliefs must exist at the time one's conclusion-belief is formed, since inferential force is conferred upon each proposition in the ordered set all at once, such that my premise-belief(s) and conclusion-belief are formed all at once.[26] On another account, my conclusion-belief is a conclusion-belief because it possesses the *form* of a conclusion-belief, and it possesses this form if and only if it stands in a formal relation that encodes taking one's premise-beliefs to support it (Kietzmann 2017, p. 300). Crucially, if I cease to have my premise-belief(s) before I form my would-be conclusion-belief, then it does not take on the form of a conclusion-belief at all, and so I have not actually inferred anything.

On either of these accounts of the Taking Condition, it seems to be a necessary condition on inference that my premise-beliefs do not extinguish before my conclusion-belief is formed.

[26] In fact, if inference begins and ends with a single act of attaching inferential force, this may overturn the orthodox view that inference is a temporally extended psychological transition from one mental state to another. This would vindicate a suggestion, put to me by an anonymous reviewer, that it is possible to perform 'simultaneous (BEL)-inferences' inferences in which one believes *p* and *I believe that p* in a single reflective temporal window. In that event, the first horn of my dilemma may be circumvented. This only works, however, if *all* (BEL)-inferences are like this. Moreover, as I will argue, below, this argument for the simultaneity of (BEL) inferences will thrust ITM onto the second horn of my dilemma.

And yet, while appealing to the taking condition on inference may rule out the possibility of inferential belief extinction, this move simultaneously undermines ITM. One problem is that, on many conceptions of the Taking Condition, inference turns out to be a self-conscious activity wherein an agent must already be aware of her premise-beliefs, such that she can take their propositional contents to support her conclusion (Kietzmann 2017; Boghossian 2014). If this is the case, then ITM cannot explain how we acquire self-knowledge of the premise-beliefs in a (BEL) inference, for one must already know that one believes P before one can take one's P-belief to support Q (where Q, in a (BEL)-inference, is a self-belief).

One might push back by arguing that taking one's premise(s) to support one's conclusion can be done merely by focusing on the propositional contents of one's mental states and the putative epistemic support relations between these, without also requiring awareness of oneself as believing these propositions.[27] If this is possible, no self-knowledge is presupposed. The problem now, however, is that the plausibility of this 'deflationary' view of takings does not generalize all the way to (BEL)-inferences. To see this, consider a (BEL)-inference in which I take the proposition *there are at least fifty sports fans in this bar right now* to epistemically support the proposition *I believe that there are at least fifty sports fans in this bar right now*. Our question now is: what epistemic support relation do I appreciate between these propositions? As observed by both Boyle and Byrne (see §III), it cannot be that I appreciate the former proposition as providing deductive or inductive support for the latter. To my mind, then, it can only be that I take the former to epistemically support the latter because believing the latter on the basis of the former is self-verifying (as Byrne argues). But how can I appreciate *this* except by taking it that the former's *being the content of my belief* is what makes its self-ascription self-verifying? I see

[27] Compare Peacocke's discussion of "'second-tier' thought" (1996, p. 129).

no other way for this to make sense, but this means that there is no way for me to take the former to support the latter without already having self-knowledge of my belief.[28]

So we reach the second horn of my dilemma: avoiding the threat of inferential belief extinction seems to render ITM explanatorily circular. I conclude that ITM fails to answer the supplementation question for Burgean Agentialists: either ITM can yield warranted yet false self-beliefs, in which case it violates a constitutive norm of critical reasoning, or it presupposes the very self-knowledge it seeks to explain.

§V—Conclusion:

I have argued that ITM does not answer the supplementation question even when we focus on the case of judgement-based belief. The importance of this result for proponents of ITM surely depends on how seriously they should take Burgean Agentialism. But Burgean Agentialism is a sophisticated picture of the epistemic value and structure of self-reflective agency, agency that many take as constitutive of developed human rationality. So the issue is hardly a small one.

No matter how we feel about this result, however, there is another worry that I want to raise at the close of this paper. To see it, note first that while not all rule following is conscious or self-conscious, it is usually relatively easy for the cognitively mature among us to recognize the reasonability of the rules we do in fact follow in different theoretical or practical domains, at least when those rules are explicitly articulated to us by someone else. Even semantic rules—plausible candidates, in many instances, for rules we learn to follow without ever first doing so self-consciously (Boghossian 2015, p. 8)—can be reflected on more or less effortlessly by ordinary agents, without one's suddenly finding them unreasonable.

[28] See Boyle (2011, p. 231) for a similar point.

The problem is that grasping the reasonability of rules like (BEL) seems to require knowledge of sophisticated philosophical arguments like those provided by Byrne. These arguments are sophisticated to the extent that they require us to recognize the rationality of (BEL)-inferences as a product of their self-verifying status, despite the fact that they flout basic deductive and inductive logical rules. I worry that many (if not most) ordinary, cognitively mature agents have not thought—nor would immediately recognize when told—that following (BEL) is reasonable in this way despite its logical shortcomings.

What this means is that, for most ordinary agents, it is implausible that they will become conscious of the fact that they follow (BEL) without—in all likelihood, at least—finding it implausible that they do so, or at least being unsure that they do so. And this strikes me as (admittedly defeasible) evidence that ordinary agents do not follow it: if they genuinely followed it, suggesting this rule to them would not strike them this way. The same, I propose, is true for Byrne's other transparent inference schemas. This is all contra Byrne, who thinks that ordinary agents use (BEL), as well as other transparent inference rules, all the time in their daily lives.

Note that the problem remains even if we reject the Taking Condition. For, while one could disagree that drawing an inference necessarily requires one to be cognizant of epistemic support relations between its premise(s) and conclusion, this is besides the point here. Rather, my claim is that, no matter whether agents necessarily take their premise(s) to epistemically support their conclusions in the course of drawing an inference, an ordinary agent is likely to balk at (BEL) when it is suggested to them as an inference rule. Note also, and finally, that I am not denying Byrne the opportunity to say—as he does to Boyle—that we follow (BEL) sub-personally. For I am not suggesting that, *as one self-ascribes* a mental state first-personally, the method that one uses to do so (and its source of warrant) must be transparent to one. Again, point is only about

what might happen when agents to become aware of all this. Even if we follow rules sub-personally, if we are really following them it seems odd to suggest that they should strike us as bizarre rules once they are made available at the person-level.

References:

Armstrong, D. 1968. *A Materialist Theory of Mind*. London/New York: Routledge.
Bar-On, D. 2004. *Speaking My Mind: Expression and Self-Knowledge*. New York: Oxford University Press.
Bilgrami, A. 2006. *Self-Knowledge and Resentment*. Cambridge, MA: Harvard University Press.
Bilgrami, A. 2012. The Unique Status of Self-Knowledge. In Coliva A. (Ed.), *The Self and Self-Knowledge*, 263-278.
Boghossian, P. 2014. What is Inference? *Philosophical Studies*, 169, 1-18.
Boghossian, P. 2015. Reasoning and Reflection: A Reply to Kornblith. *Analysis*, 76(1): 41-54.
Borgoni, C. 2015. On Knowing One's Resistant Beliefs. *Philosophical Explorations*, 18(2): 212-225.
Borgoni, C. & Luthra, Y. 2017. Epistemic Akrasia and the Fallibility of Critical Reasoning. *Philosophical Studies*, 174: 877-886.
Borgoni, C. 2018. Basic Self-Knowledge and Transparency. *Synthese*, 195: 679-696.
Boyle, M. 2011. Transparent Self-Knowledge. *Proceedings of the Aristotelian Society, Supplementary Volumes,* 85, pp. 223-241.
Burge, T. 1996. Our Entitlement to Self-Knowledge. *Proceedings of the Aristotelian Society*, 96, 1-26.
Burge, T. 2013. "Self and Self-Understanding: The Dewey Lectures (2007, 2011)" in T. Burge (Ed.), *Cognition Through Understanding: Philosophical Essays, Volume 3*. Oxford: Oxford University Press, pp. 140-226.
Byrne, A. 2005. Introspection. *Philosophical Topics*, 33(1), 79-104.
Byrne, A. 2011a. "Knowing What I Want" In J. Liu & J. Perry (Eds.) *Consciousness and the Self: New Essays*. Cambridge, UK: Cambridge University Press.
Byrne, A. 2011b. Transparency, Belief, Intention. *Aristotelian Society Supplementary Volume*, 85, 201-221.
Byrne, A. 2018. *Transparency and Self-Knowledge.* Oxford University Press.
Cassam, Q. 2015. *Self-Knowledge for Humans*. Oxford: Oxford University Press.
Coliva, A. 2009. Self-Knowledge and Commitments. *Synthese*, 171, 365-375.
Coliva, A. 2012. "One Variety of Self-Knowledge: Constitutivism as Constructivism." In Coliva, A. (Ed.) *The Self and Self-Knowledge*, 212-242. Oxford University Press.
Coliva, A. 2016. *The Varieties of Self-Knowledge*. London, England: Palgrave Macmillan.
Davidson, D. 1984. "First Person Authority" *Dialectica* 38.2-3 (1984): 101-112.
Evans, G. 1982. *The Varieties of Reference*. Oxford University Press.
Finkelstein, D. 2003. *Expression and the Inner*. Harvard University Press.
Gertler, B. 2011a. *Self-Knowledge*. Routledge.
Gertler, B. 2011b. "Self-Knowledge and the Transparency of Belief" in A. Hatzimoysis (Ed.), *Self-Knowledge*, 125-145. Oxford University Press.

Gertler, B. 2016b. Self-Knowledge and Rational Agency: A Defense of Empiricism. *Philosophy and Phenomenological Research*, 96(1): 91-109.
Hlobil, U. 2019. Inferring by Attaching Force. *Australasian Journal of Philosophy*, https://doi.org/10.1080/00048402.2018.1564060
Komorowska-Mach, J. 2019. Introspection—One Or More? Pluralism about Self-Knowledge. *Filozofia Nauki* 1(105): 5-25.
Kietzmann, C. 2017. Inference and the Taking Condition. *Ratio*, DOI: 10.1111/rati.12195
Leite, A. 2018. Changing One's Mind: Self-Conscious Belief and Rational Endorsement. *Philosophy and Phenomenological Research, 97(1)*, 150-171.
Lycan, W. 1996. *Consciousness and Experience*. Cambridge, MA: MIT Press.
Owens, D. 2000. *Rationality without Freedom*. London: Routledge.
Owens, D. 2011. "Deliberation and the First-Person." In Hatzimoysis A. (Ed), *Self-Knowledge* (pp. 261-278), Oxford University Press.
Parent, T. 2017. *Self-Reflection for the Opaque Mind: An Essay in Neo-Sellarsian Philosophy*. Routledge.
Peacocke, C. 1996. Our Entitlement to Self-Knowledge: Entitlement, Self-Knowledge and Conceptual Redeployment. *Proceedings of the Arisototelian Society, 96(1)*, 117-58.
Peacocke, C. 1998. Conscious Attitudes, Attention, and Self-Knowledge. In Wright, C., Smith, B. & Macdonald, C. (Eds.), *Knowing Our Own Minds*, 63-98. Oxford University Press.
Peacocke, A. 2017. Embedded Mental Action in Self-Attribution of Belief. *Philosophical Studies*, 170: 353-377.
Peters, E. 2017. Introspection, Mindreading, and the Transparency of Belief. *European Journal of Philosophy*, DOI: 10.1111/ejop.12318
Reed, B. 2010. Self-Knowledge and Rationality. *Philosophy and Phenomenological Research*, 80(1): 164-181.
Samoilova, K. 2016. Transparency and Introspective Justification. *Synthese*, 193: 3363-3381.
Shoemaker, S. 1996a. *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.
Sorgiovanni, B. 2019. The Agential Point of View. *Pacific Philosophical Quarterly*, 100(2): 549-572.
Williamson, T. 2000. Knowledge and its Limits. Oxford: Oxford University Press.
Wright, C. 2001. *Rails to Infinity*. Harvard University Press.
Zimmerman, A. 2006. Basic Self-Knowledge: Answering Peacocke's Criticisms of Constitutivism. *Philosophical Studies*, 128: 337-379.