Ethical Theory and Moral Practice (2021) 24:65–86
https://doi.org/10.1007/s10677-020-10118-8

The Value of Transparent Self-Knowledge

Fleur Jongepier[1]

Accepted: 11 August 2020 / Published online: 9 December 2020

Abstract
Questions about the normative significance of 'transparency' do not receive much attention, even though they were central to Richard Moran's (2001) original account. Instead, transparency is typically studied because of its epistemic and psychological peculiarities. In this paper, I consider three normative conceptions of transparency: teleological rationalism, procedural rationalism, and relational rationalism. The first is a theory about how transparency might relate to flourishing as a rational agent; the latter two are theories about how transparency relates to non-alienated self-knowledge. All three conceptions, I argue, face serious problems. I end the paper by suggesting where the rationalist might go from here and by highlighting the importance of focusing not on the methods but on the broader circumstances in which self-knowledge is gained or lost.

Keywords Alienation · Authenticity · Autonomy · Self-knowledge · Transparency

Introduction

The notion of transparency relevant to the contemporary self-knowledge debate is typically explained by invoking the metaphor that instead of looking inwards, one acquires self-knowledge by looking outwards. Mental states are not facts of the matter waiting to be discovered with some sort of inner scanner. Instead, one acquires knowledge of one's own mental states by actively reflecting on the proposition one's attitudes are about, or so transparency theorists claim. One acquires self-knowledge by treating questions about one's mental states, like "Do I believe that a third world war is coming?" as *transparent to* corresponding world-directed questions, such as, "Is it true that a third world war is coming?" (Evans 1982). When confronted with self-knowledge questions, one 'looks through' one's attitudes, directly at the object one's attitude is about.

✉ Fleur Jongepier
f.jongepier@ftr.ru.nl

[1] Faculty of Philosophy, Theology and Religious Studies, Radboud University Nijmegen, Nijmegen, the Netherlands

One of the main attractions of transparency views is the idea that self-knowledge and agency are intimately connected, and that traditional introspectionist and interpretationalist accounts of self-knowledge have overlooked this important connection.[1] Arguably the most influential transparency account to date is the one developed by Richard Moran (2001).[2] Specific to Moran's approach is that the capacity to answer world-directed questions – or, as Moran calls it, 'obeying' or 'conforming to' transparency – is intimately connected to one's capacity for rational deliberation.

Acquiring transparent self-knowledge, on Moran's view, involves 'making up one's mind' which involves rationally reflecting on, and coming to a judgment about, states of affairs in the world.[3] For instance, in order to know whether I believe that a third world war is likely to happen, I actively reflect on, say, the current situation in the US, Russia, Hungary, and so on. After weighing the reasons pro and con, I come to a judgment which enables me to know what my belief is.[4] The transparency method is meant to allow one to acquire knowledge not just of one's beliefs, but of any "judgment-sensitive" attitude, including one's desire to "change jobs, or learn French, or avoid being seen" (Moran 2001, 115). Following Brie Gertler (2011) and Quassim Cassam (2014), I will use the label 'rationalism' to refer to the view according to which self-knowledge is acquired by making up one's mind, the latter of which involves rational deliberation.

Over the years, the transparency approach generally, and rationalism in particular, has come in for some heavy criticism. Critics have for instance worried about rationalism's scope, and have argued rationalism cannot properly account for knowledge of our emotions and other affective states (Cassam 2014; Kloosterboer 2015; Strijbos and Jongepier 2018). It has also been claimed that in some situations, making up one's mind in order to acquire self-knowledge leads to self-deception and estrangement (Lear 2004), and that following the transparency method may even lead to "barking up the wrong tree" (Cassam 2014, 105).

The most commonly voiced critique however is that having to actively engage in world-directed deliberation is an overly intellectualistic requirement on self-knowledge. Making up one's mind is not the normal or routine way of acquiring self-knowledge, and the procedure is said to simply be too demanding, sometimes even counterproductive (Byrne 2018; Carman 2003; Finkelstein 2012; McGeer 2007; Shah and Velleman 2005; Shoemaker 2003). As Cassam puts it, rationalism presents a "highly unrealistic conception of human self-knowledge" and is in urgent need of a "reality check" (Cassam 2014, 52). What these objections all have in common is that they are epistemically-psychologically oriented objections about transparency being an overly demanding or ineffective method of self-knowledge. I'll refer to this cluster of objections about the adequacy of the transparency method as constituting the 'standard objections' against rationalism.

[1] Introspectionism is the view that self-knowledge is acquire by some form of inner perception; interpretationalism is the view that self-knowledge is acquired through an (implicit or unconscious) process of interpretation or by turning one's general folk psychological capacities upon oneself.

[2] Earlier transparency approaches include Tugendhat (1986) and McGeer (1996).

[3] A brief note on terminology. When using the phrase 'making up one's mind' I shall for sake of simplicity take this to be equivalent to 'obeying/conforming to the transparency condition' and 'following the transparency method/procedure'. Distinctions can and perhaps should be made here, but they are not important for the arguments in this paper.

[4] How one gets from (knowledge of) one's judgments to (knowledge of) one's beliefs is a question I cannot go into here.

If, however, transparency is meant to provide a "normative demand" or "normative requirement" (Moran 2001, xvi–xvii), then the fact that rationalism does not provide an empirically realistic account of self-knowledge should be no immediate cause for concern. It may well be true that we are often unable or unwilling to 'conform' to transparency or that doing so is costly. The normative point would be that we nonetheless should.

According to critics of rationalism, the preferred method and thus theory of self-knowledge should be adjusted to our epistemic and psychological limitations. But if obeying transparency is valuable or normatively significant in some important way, then epistemically and psychologically limited human beings should adjust themselves to the theory, not vice versa. Acquiring self-knowledge by turning our folk psychological capacities upon ourselves or by employing self-tracking devices or conducting psychological self-tests might be a much more efficient method of self-knowledge, but a normative rationalist would say that acquiring transparent self-knowledge might nonetheless be more valuable. So, if rationalism is first and foremost in the business of making normative claims, it may be resistant to epistemic-psychological objections. Rationalism would be, as I'll call it, 'normatively immune' to standard objections.

Most critics of rationalism have not been particularly interested in whether and how transparency presents a normative demand exactly, as a result of which the rationalist's normative claims (about freedom, rational agency, well-being, and alienation) have not been subjected to critical scrutiny in any serious way.[5] This paper aims to pave the way for evaluating rationalism's distinctly normative claims by first of all getting clear on what those claims might be.

I shall proceed as follows. In the next section, I first of all say a bit more about what might make rationalism a normative theory of self-knowledge and offer a diagnosis on why the rationalist's normative claims have so far been overlooked. In section 2, I take my cue from Boyle's work and reconstruct what I call 'teleological rationalism'. On this view, the capacity for making up one's mind is essential to rational agency and is necessary for flourishing or living a worthwhile life. I argue this approach ends up in a dilemma: it either makes too strong claims and ends saying that inferentialists and self-trackers can't flourish, or it makes too weak claims and transparency is no longer normatively significant in a way that makes rationalist immune to standard objections. In section 3, I go on to consider the view according to which obeying the transparency condition is necessary and sufficient for acquiring non-alienated self-knowledge. I argue this approach has more normative potential, but runs into what I will call the 'false positive-objection' (section 4). In section 5, I consider a way of avoiding the false positive objection by revising normative rationalism along 'relational' lines, just as feminist theorists have done for the concept of personal autonomy. In the final section, I argue that this new relational version of normative rationalism has rich resources so far missing in the self-knowledge debate, but suffers from some serious methodological problems. The upshot is mostly negative for all versions of normative rationalism. I end in a somewhat more constructive vein, by suggesting that relational rationalism may point towards a new, particularist way

[5] A notable exception is Cassam (2014, chap. 7), to which this paper is much indebted. Cassam offers a distinct conception of normative rationalism. He reconstructs it as the view that we are supposed to, or ought to, "approximate" what he refers to as "homo philosophicus", that is, a "model epistemic citizen" whose "attitudes are as they ought rationally to be" (Cassam 2014, 83). Getting normativity from ideal-conceptions is indeed one way of getting a normative foothold, but it's not the only way, and it doesn't seem to be the route rationalists themselves want to take (cf. Boyle 2015). Other authors who have engaged with rationalism's normative claims include Victoria McGeer (2007), Carla Bagnoli (2007), and Jonathan Lear (2004).

of thinking about self-knowledge: one that is less preoccupied with the various methods of self-knowledge and their differences, and more with the bodily, emotional, and socio-political circumstances that make these methods (in)effective in the first place.

1 The Contours of Normative Rationalism and a Diagnosis

Standard objections to rationalism often presuppose a specific take on transparency. First, critics implicitly or explicitly take transparency to be a 'method' or 'procedure' (Borgoni forthcoming; Byrne 2018; Cassam 2014, 2015; Finkelstein 2012; Golob 2015). Second, they take rationalists to be (primarily) concerned with making claims about our psychologies and epistemic strategies. Third, rationalist's claims about 'immediacy' are often taken to be the most important and/or controversial aspect of rationalism. Finally, critics often assume that normative claims about e.g. alienation can bracketed when assessing the epistemic and psychological strengths of transparency as a method of self-knowledge. In short: critics take rationalists to claim that 'following the transparency method' is the only or in any case the best way towards acquiring non-inferential self-knowledge.

Interestingly, rationalists themselves never talk about transparency as a method or procedure. Instead, they prefer to talk about the transparency 'condition'. On the face of it, a method or procedure is a psychological-epistemic notion that refers to something one might but need not follow, whereas a condition is something one ought to meet or obey. This difference could be evidence of the fact that epistemically oriented critics of rationalism have overlooked the distinctively normative spirit of rationalism.[6]

Moran himself explicitly positions his theory of self-knowledge in the context of normative concepts such as freedom, autonomy, and alienation, and mentions that transparency is meant to provide a "normative demand" or "normative requirement" (Moran 2001, xvi–xvii). He writes that transparency has a "deeper relation to freedom or rationality than any other one (e.g., various modes of perception) and that transparency matters to the "overall psychic health" and "well-being of the person" (2001, xxvi; 136–37). He also claims that (not) obeying the transparency condition can explain the meaning of some of our moral experiences and attitudes, and helps getting a better understanding of why "in general, a person's attitude toward himself makes a difference to how we feel about him" (2001, 183). One of the principal aims of *Authority and Estrangement*, he writes, is "to make a start at showing how some of these seemingly remote matters from philosophy of mind have a genuine role to play in accounting for aspects of the structure and phenomenology of moral experience" (2001, 193). Recently, Matthew Boyle (2015) also calls attention to the normative aspects of rationalism, by claiming that transparency is connected to the Sartrean notion of "bad faith".

An insightful example that clearly shows Moran's normative ambitions is the following: imagine asking someone whether she intends to pay back the money she borrowed. Suppose she answers, "As far as I can tell, yes" (2001, 26). Moran's point is clearly not that giving what Moran would call a 'theoretical' or 'third-personal' (i.e. non-transparent) answer wouldn't constitute an answer to the question. Rather, the point is that that would be a *strange* answer or that we shouldn't be prepared to *accept* such an answer. This example makes clear that at least

[6] This terminological difference might in the end not amount to much. And even if transparency *were* a method, it could still be a valuable method to follow. In any case, whenever I talk about transparency as a method or procedure, I shall take it to be a method that has (prudential or moral) value. More on 'value' below.

for Moran, the rationalist's concerns are not primarily psychological or epistemic, but normative.

What exactly we should take transparency understood as a "normative demand" to be, is the very ambition of this paper to attempt to clarify. But, minimally, I propose to understand any normative conception of transparency as claiming that transparency is something (1) one ought to follow or obey; and (2) that transparency has 'value' and is somehow worth obeying, having, or striving towards.[7]

Some terminological remarks are in order. First, rather than understanding value in a narrow sense as implying a value theory/axiology, I shall understand the notion in a deliberately broad fashion, covering both transparency's possible relation to the good and bad as well as to the deontic concepts of right and wrong. Second, other than epistemic value, transparency may have prudential or moral value, or both (or neither). Claims about the relation transparency bears to mental health and wellbeing would exemplify the possible prudential value of transparency, while claims about how transparency matters to our reactive attitudes, freedom, or (not) treating oneself as a mere means, would on the face of it illustrate transparency having moral value. Third, wherever I talk of transparency being 'normatively significant,' I shall have in mind that it 'has value' in the above sense.

One might wonder, if it's indeed the case that critics have paid insufficient attention to rationalism's distinctly normative claims, why this might be so. The diagnosis, I believe, is two-fold. First of all, critics of rationalism are usually not moral philosophers but epistemologists and/or philosophers of mind. Hence, they are typically interested in knowledge generally, and study self-knowledge as a special case. As Cassam points out: "many philosophers of self-knowledge have concentrated on the epistemology of relatively trivial self-knowledge" (2014, 11). Fascination with transparency is not due to the fact that transparency matters to authenticity, autonomy, or something in the normative domain, but to the fact that transparency is somehow theoretically puzzling. How could one possibly come to know that one believes that it is raining by checking on the rain? After all, "meteorology sheds little light on psychology!" (Byrne 2018, 4). The main aim has been to dispel the puzzle.

A second possible explanation is that the lack of attention to normative questions is due to rationalists themselves. Moran for instance remains largely silent on the question of what he takes the concepts of 'mental health,' 'autonomy' or 'rational freedom' to be, and does not compare or contrast his use of these notions with the way in which these concepts are typically understood in other debates.[8] Instead, Moran and other rationalists have been mostly concerned with arguing against introspectionist and interpretationalist views. Moreover, rationalists have devoted a lot of time and energy to defending the thesis that transparent self-knowledge is an immediate, non-inferential type of self-knowledge, which many epistemologists and philosophers of mind on their turn have found controversial. This, in all likelihood, has resulted in the fact that many theorists have subsequently focussed on the rationalist's psychological-epistemic claims regarding transparency and have ignored its connection to normative notions.

It's time to ask: in what way might we understand transparency as something that has value or is worth striving towards?

[7] The second claim is necessary because the first may 'merely' include epistemic norms or constitutive conditions. This will become relevant when discussing teleological rationalism.

[8] One exception is Moran (2002).

2 Conforming to Transparency as an Essential Capacity

2.1 The Contours of Teleological Rationalism

On different occasions, Moran gives the impression that what he is after is to provide a theory about how transparency is related to being a rational agent. In the beginning of his book, Moran for instance describes his project as one of "trying to do justice to a certain tension in our thinking about the *possibilities* of self-knowledge" and "the distinctiveness of the first-person perspective more generally" (2001, xxx). Elsewhere, he writes that what is central about transparency is not that one *normally* arrives at one's beliefs by asking world-directed questions, but merely that "there is logical room for such a question" (2001, 63). In response to the objection that his approach is hyper-intellectualistic, he stresses that his aim was only to make a "modest claim", "something the denial of which would be equivalent to denying that people ever actually reason to a conclusion" (2004, 458). It is not entirely clear, though, what the status of claims like these is exactly. Happily, in one of his papers, Matthew Boyle sets it has his "principal aim" to "clarify" the rationalist conception about "the nature of belief and judgment and how concepts of agency relate to them" (Boyle 2011, 4).

Rational creatures, Boyle writes, "are distinguished by their capacity for a special sort of cognitive and practical self-determination, a capacity which makes their relation to their own mental lives fundamentally different from that of a nonrational animal" (2011, 1). Clearly, for Boyle, 'transparency' is not first and foremost a psychological 'procedure' (effective or ineffective) but rather an essential capacity. Or, as he typically puts it, a "power" that is distinctive of the sort of beings we essentially are. Elsewhere, Boyle writes: "I want to understand what sort of distinction writers in the Aristotelian tradition meant to be drawing when they distinguished rational from nonrational minds, and what sort of depth they were claiming for this distinction" (Boyle 2015, 346). What would make such claims normatively interesting, though? We need an answer to this question if rationalism is to be normatively immune to standard objections.

Here we can consider some suggestions by other Aristotelean thinkers. As Martha Nussbaum writes, when we speak of our human nature we are considering aspects that are "in the broadest sense, ethical", because we are considering beliefs "about what is worthwhile and worthless, liveable and not liveable" (Nussbaum 1995, 101). A natural interpretation (not explicitly considered by Boyle but which would make sense the references to Aristotle), would be to understand Moran's and Boyle's claims as part of what Julia Driver refers to as "the well-functioning view", that is, the view that "a creature flourishes when it functions well" where the definition of functioning well involves determining the functions that are distinctive of the creature (Driver 2001, 96). In order for teleological rationalism to be sufficiently normative, that is, normative enough for rationalism to be immune to standard objections, we might take the teleological rationalist to claim that a distinctively rational creature would only flourish, function well or lead a life that's worthwhile, if it has the capacity to "obey transparency".

2.2 Evaluating Teleological Rationalism

Teleological rationalism, as I have briefly reconstructed it, would have some obvious advantages. The view first of all makes exegetical sense. It makes sense of the claims that rationalists like Moran and Boyle have made about how transparency is related to deliberative agency and our rational nature. A more important advantage is that the view appears, on the face of it, to be

able to claim normative immunity. After all, saying a capacity is essential to human flourishing is compatible with the idea that it is not a capacity we routinely exercise, or that it is not the most effective way of acquiring self-knowledge. In fact, it's in principle compatible with our never doing so at all. The normative claim is that we either ought to *have* the capacity or ought to *exercise* it more often (this difference will become important in a moment).

So far so good. However, there are two possible objections which call into question teleological rationalism's true normative potential. One might first of all think the teleological rationalist's "specification of the essence of humankind" (Boyle 2012, 399) is inaccurate. Cassam for instance points out that if psychologists and behavioural economics are right that human beings often make irrational decisions, act on 'auto-pilot', suffer from self-ignorance, implicit biases and all the rest, then that makes it hard to keep insisting that our true essence lies in our capacity for practical reasoning or transparent question-settling.[9]

A second but related worry is that teleological rationalism may be insufficiently considerate about, if not to say disrespectful towards, individuals who engage in non-transparent varieties of self-understanding. If teleological rationalism makes normative claims about what is valuable or worthwhile (and it's not clear how else one might claim normative immunity) then this would appear to have the uncomfortable implication that those who are either unable or unwilling to 'make up their minds' would be leading less worthwhile or valuable lives.

Rather than discussing the possible ways for the teleological rationalist to reply to these objections and evaluating those replies, I want to highlight instead the dialectic that will inevitably occur between rationalists and their critics (inferentialists, mostly). In responding to objections like those above, the teleological rationalist will have to weaken her claims, as a result of which normative immunity against standard objections is likely to be lost. Consider the objection that our true nature may not be 'rational' at all. The teleological rationalist will in all likelihood stress that they only meant to make a very "modest claim" (cf. Moran 2004, 458), and will point out that their claims about our true essence were meant to be psychologically non-committal and perfectly compatible with all (replicable) results from social psychology and behavioural economics. Nussbaum for instance explicitly mentions that her account about the significance of practical reasoning "excludes only people who live without planning and organizing their lives at all: the sort of creature, we might say, who would be the survivor of a frontal lobotomy" (Nussbaum 1995, 117). Rationalists might similarly say that only survivors of a frontal lobotomy are unable to conform to "transparency as a normative demand" (Moran 2001, xvi).

The good news is that pretty much everyone can meet these demands. The bad news is that, ironically, meeting the transparency condition turns out not to be too demanding, as critics originally claimed, but not demanding enough. If everyone except for people suffering from frontal lobotomies can meet the demand whatever they do, then it's hard to see what value 'obeying of the transparency condition' has, since it's not clear how one might fail to obey it.[10] Also, it's a bit of a stretch to say that the sort of creature who survived a frontal lobotomies (or creatures who just don't happen to have the relevant capacities, like rabbits or robots) would be 'irrational' or that there would be value *for them* if they were capable of making up their minds, if indeed they had minds.

[9] Cf. Cassam, *Self-Knowledge for Humans*, 78ff.

[10] Cf. Korsgaard (2008, 6–7): "A principle that moves us inevitably cannot serve as a guide, for it is not possible to be guided unless it is also possible to fail to be guided."

Again, there will be ways – good ways – for rationalists to respond. But notice that againthe teleological rationalist is likely to stress the modesty of their claims. By saying that obeying transparency is "valuable" for creatures or that it is important to living a flourishing life, they meant something weaker than our folk notion of value or flourishing. In fact the rationalist will probably want to drop the claims about flourishing and living worthwhile lives at this point. Instead what we should say is simply that the relevant capacity is inevitable: it is a necessary condition for being rational or minded at all. A normative rationalist will thus probably retreat to a transcendental or constitutive rather than a teleological claim.[11]

However, this transcendental/constitutive claim is – in spite of being much more ambitious, metaphysically speaking[12] – at the same time normatively speaking potentially rather underwhelming. In any case it looks like the rationalist can no longer claim normative immunity from standard objections. Imagine an inferentialist questioning the transparency method by pointing out we hardly ever make up our minds in order to acquire self-knowledge, and the rationalist responds by saying 'perhaps so, but without the capacity we would not be rational'. That would be a bit disappointing. Consider also Moran who emphasizes that "what is essential ... is that there is *logical room* for [deliberative] questions" (2001, 63). Taylor Carman rightly responds by saying that "it is important to recognize how weak that claim is", since it tells us only that "one can raise them without threat of inconsistency" and it "says nothing about the *relevance or propriety* of such questions" (Carman 2003, 404 emphases added). Relatedly, John Christman remarks that "a mere *capacity* to reflect is too weak" because a person might well have "a capacity to reflect on herself", or answer world-directed questions, "but never does" (Christman 2005, 334). The rationalist needs more than just logical rooms and certain deliberative powers to be immune to objections.

To be sure, none of this makes the teleological project trivial or worthless by any means. Boyle has done more than anyone in working out a detailed account of rational agency and offering a novel and intriguing metaphysics of mind. The worry in the present context is that an intriguing metaphysics of mind does not make rationalism, as such, sufficiently normatively robust. The dialectic sketched above leads to the following dilemma for the teleological rationalist: either teleological rationalism is too weak in that the transparency condition is ironically only worth obeying for creatures who suffered a frontal lobotomy, or else it's too strong and ends telling people who engage in non-transparent methods of self-knowledges as not properly flourishing.

There is one way of dealing with this dilemma, which is to focus not on the value of having a *capacity* to make up one's mind and answer world-directed questions, but rather on the value of actually *exercising* that capacity, and to claim that not exercising the capacity on specific occasions is (normatively) problematic. It is unclear which claim rationalists are actually after – hence the main aim here is meant to lay out the theoretical options – but there is some exegetical evidence for the "exercise" view. For instance, in recent work Boyle writes: "Where transparency does not obtain, we can say that a person is *alienated* from her own belief: she is not capable of knowing it from participant's standpoint, so to speak" (Boyle 2015, 341). Here,

[11] Some may want to argue this is a false dichotomy. After all, transcendental, teleological and constitutivist claims and principles come in many shapes and sizes (Haase and Mayr 2019). Perhaps it is a false dichotomy, perhaps it's not. The argumentative strategy pursued here is an attempt to shift the burden of proof and to stress that it's just not clear which claim rationalists are making. If it's a metaphysical-teleological claim – which I believe in the end is likely to be Boyle's strategy – that's fine. But then one cannot claim normative immunity from standard objections.

[12] For a discussion of the "formidable difficulties" of transcendental arguments, see e.g. Cassam (1987).

Boyle is talking not about powers but about cases in which the relevant capacity was not exercised, and *that's* what's problematic. And this certainly looks like a substantially normative claim, assuming at least that alienation is never a good thing.[13]

But now a new problem arises, for we must now ask how the "exercise" claim relates to the "capacity" claim. Are you alienated with respect to your essential rational nature if you frequently fail to make up your mind, or are you already alienated if you fail to conform to transparency just once (as Boyle seems to suggest)? Do you suffer from alienation if you perform an implicit association test, infer your emotions from your diary, or if you take a friend's or therapist's testimony about what they think you want as good enough evidence for what you in fact want? In each of these cases, you still have the power to make up your mind, i.e. to obey transparency. It's just that you are unwilling or unable to do so on a specific occasion.

These are difficult questions, and normative rationalists will no doubt have interesting things to say in response to them. But notice that we have now moved from claims about the significance of there being "logical room" for asking certain questions and the inevitability of having a certain capacity, to claims about the value of exercising that capacity on specific occasions. But these are distinct claims, and the one doesn't follow automatically from the other. One might after all say that what defines us is a capacity for transparent question-settling and yet say that frequently exercising that capacity is prudentially or morally speaking a bad idea. The two claims may well be connected in interesting ways, but they are distinct, and must be evaluated separately. I therefore now turn to this second conception of normative rationalism, which involves a claim about actually exercising our capacity to make up our minds.

3 Transparency as Non-alienation

The second normative conception of transparency as I will reconstruct it consists of two claims: first, obeying transparency provides the subject with a distinct type of self-knowledge, namely, 'non-alienated self-knowledge'.[14] Second, having such self-knowledge is valuable in some way. What I plan to do in this section and the next is explain what non-alienated self-knowledge would be according to rationalism, why having such knowledge would matter, and whether this normative conception of rationalism fares any better than its teleological cousin.

In *Authority and Estrangement*, Moran typically explains transparency negatively by contrasting it with non-transparent ways of acquiring self-knowledge. He interchangeably refers to these as 'empirical', 'theoretical', 'third-personal' or 'impersonal' ways toward self-knowledge.[15] To better understand why, according to Moran, non-transparent self-knowledge amounts to alienated self-knowledge, consider his central example of a person undergoing psychotherapy:

[13] A debatable assumption, but one that I shall not debate here.

[14] One could also refer to this as "authentic self-knowledge" or "autonomous self-knowledge". Much has been written on the distinction between authenticity and autonomy, and distinguishing between authentic and autonomous self-knowledge may be useful. But given that Moran does not distinguish between non-alienation, authenticity, and autonomy, I will likewise treat them as equivalent for present purposes.

[15] These are not in fact equivalent. See Strijbos and Jongepier (2018).

> The person who feels anger at the dead parent for having abandoned her, or who feels betrayed or deprived of something by another child, may only know of this attitude through the eliciting and interpreting of evidence of various kinds. She might become thoroughly convinced, both from the constructions of the analyst, as well as from her own appreciation of the evidence, that this attitude must indeed be attributed to her. And yet, at the same time, when she reflects on the world-directed question itself, whether she has indeed been betrayed by this person, she may find that the answer is no or can't be settled one way or the other. So, transparency fails because she cannot learn of this attitude of hers by reflection on the object of that attitude. She can only learn of it in a fully theoretical manner, taking an empirical stance toward herself as a particular psychological subject. (2001, 85)[16]

What makes the person undergoing psychotherapy alienated from her own mental states, according to Moran, is that she is unable to consider the objects or contents of her states directly. She considers her mental states *as* mental states and herself not as an agent but as a psychological subject like any other. The alienated self-perspective is thus defined in terms of a perspective that does not look through one's attitudes but *at* them, or *at* oneself. The transparent, non-alienated self-perspective on the other hand is one that "involves no essential reference to oneself at all" (2001, xvi).

The language of alienation brings out the fact that although alienated self-knowledge is also a type of self-knowledge, it is not the 'right' or 'proper' type of self-knowledge, and that transparent self-knowledge is the good or healthy variety. This is evident from Moran's claims that there's something "wrong" with the person who can only rely on behavioural evidence to report on her mental states (2001, 68). According to Moran, conforming to transparency on the one hand and engaging in self-interpretation or self-observation on the other are "*different routes* to knowledge of the *same thing*" (2001, 89 emphases added). This means that we should interpret claims about the importance of 'making up your mind' not principally as a claim about how we as a matter of fact do acquire self-knowledge, nor as a claim about an essential 'capacity', but as a claim about the special value of a specific sort of self-knowledge.

The non-alienation view thus appears to be able to claim normative immunity. Objections about making up one's mind not being the routine method of self-knowledge or as being ineffective are neither here nor there if there's independent value in acquiring transparent self-knowledge. If opaque methods of self-knowledge such as self-interpretation, psychotherapy, reading self-help literature or using self-tracking devices are (becoming) the routine methods of self-knowledge, then what we should say is not that we have reason to be worried about normative rationalism, rather, we should be worried about the growing influence of these non-transparent methods of self-knowledge.

But despite its potential, this version of normative rationalism faces a serious problem: transparency may not actually turn out to be a plausible criterion of what it means to have non-alienated self-knowledge.

[16] Moran only considers psychoanalysis, but elsewhere he talks about "therapeutic contexts" in general.

4 The False-Positive Objection

The problem in a nutshell is that 'transparency as non-alienation' leads to false positives, that is, cases where, according to normative rationalism, we would have to say a subject has non-alienated self-knowledge whereas intuitively she lacks such knowledge. Consider the case of Patty Hearst:

> *Patty Hearst:* On 4 February 1974, Patricia Hearst, heiress and granddaughter of the powerful US media magnate William Randolph Hearst, was kidnapped by an organization calling itself the Symbionese Liberation Army (SLA). She was kept bound and blindfolded in a closet for several weeks, physically assaulted, forced to have sex with SLA members, and threatened with death. Meanwhile the SLA demanded a ransom from the Hearst Corporation, including not only requests for money but for a food giveaway worth millions of dollars and the release of two SLA members jailed for murder. On 14 April of the same year Patty Hearst caused a sensation by participating in the SLA robbery of a bank in San Francisco, after which she publicly denounced her family and expressed her commitment to the SLA. (K. Taylor 2006, 10–13)

Though this example raises many questions, I want to zoom in on a specific one: did Patty Hearst have non-alienated knowledge of her attitudes? To see how a normative rationalist would answer this question, we would need to answer the question of whether she is capable of 'conforming to transparency'. This means we have to ask: is Patty Hearst able to deliberate about world-directed questions and come to a judgement? It's hard to see why not. For instance, it is easy to image that she is able to consider a world-directed question such as "is the SLA admirable organization?" or "is my family despicable?". She would then deliberate about the reasons pro and con and finally come to a judgement: "yes, the SLA is an admirable cause" or "yes, my family is despicable". So, we can suppose that Patty thereby makes up her mind and constitutes her attitudes. The rationalist thus seems committed to saying that Patty Hearst has made up her mind and acquired non-alienated self-knowledge in the process. She knows, in a non-alienated way, what she believes and what she wants.

But if to brainwash someone is to "change a mind radically so that its owner becomes a living puppet—a human robot—without the atrocity being visible from the outside" (Hunter 1956), and if this is what happened to Patty Hearst, then should we really say that she had non-alienated self-knowledge? This conclusion would appear to overlook what is central to the example, namely that the answers Patty Hearst gives to world-directed questions are mere echoes of the SLA. The conclusion that Patty Hearst has 'all it takes' as far as 'healthy' self-knowledge is concerned, is hard to accept. If we take transparent question-settling to be the criterion for non-alienated self-knowledge, it appears to provide us with a false positive. This objection is a version of the 'garbage in, garbage out'-objection familiar from other areas of philosophy: if you put garbage into some procedure (a deliberative procedure in this case), then one can only expect garbage to be the result.[17]

Let me add a point of clarification to avoid misunderstandings. The question that the false positive-objection is meant to raise is whether one can have non-alienated self-knowledge if one does not have what we would intuitively call 'a mind of one's own'. But this is not meant to be a rhetorical question. The main point of the false positive-objection is not meant to be that

[17] This objection is typically discussed in the context of the debate about Rawls' method of reflective equilibrium, see esp. Karen Jones (2005).

Patty Hearst *obviously* does not have non-alienated self-knowledge. The point is rather that it does not appear to be clear that she does, either. In other words, whatever we say, it should not be evident whether or not she has non-alienated self-knowledge. Any theory that gives us an easy and straightforward answer – as rationalism does, in its present form – when dealing with such difficult cases should be regarded with suspicion. So, there are actually two options for the rationalist: either to say they are genuinely confronted with a false positive and bite the bullet, or else to explain why, on reflection, it is not actually a false positive.

The problem with the second option is that the rationalist does not appear have the theoretical resources to handle hard cases like Patty Hearst's in a satisfactory manner, that is, in a way that respects its nuances and intricacies.[18] It is no coincidence, I suspect, that one of Moran's central examples involves the trivial question of whether one believes it is raining.[19] It is ironic that a theory of self-knowledge that stresses its connections to well-being, mental health, our moral attitudes, freedom and autonomy, would treat knowledge of one's belief that it is raining in the same way that it treats knowledge of someone's desire to remain faithful to those who kidnapped them.[20]

Why don't rationalists have the materials to say something more nuanced about cases like Patty Hearst? This is, I believe, no accident. A possible diagnosis is that rationalists typically want to remain neutral on what would count as *good reasons* for making up one's mind. This 'proceduralist' aspect of rationalism, as we might call it, is admittedly one of its attractions: one does not have to make any claims about whether one can have self-knowledge in circumstances which involve brainwashing, manipulation, oppression, and so on. But this proceduralist feature is also its weakness. Answering world-directed questions can, depending on the circumstances and the relevant 'deliberative material', lead to greater alienation, self-deception, and potential harm.[21]

A possible response the rationalist might give is that the false positive objection only arises because it confuses (at least) two different types of alienation. When rationalists talk about alienation, then *all they mean* is that one's relation to one's mental states is opaque. The notion of alienation that seems to be implicit in the false positive objection, on the other hand, is alienation in the sense that involves a lack of higher-order identification or feelings of repudiation (along the lines of e.g. Frankfurt 1971; Christman 1991).[22] The reason that we have the intuition that Patty Hearst is alienated may for instance be due to the fact that we are inclined to think that if she were given the chance to consider the "processes and conditions that led to the adoption" of her desires and intentions, she would resist them (Christman 1991,

[18] Arguably one of the resources that's missing here is that rationalism makes insufficient room for the diachronic nature of agency and deliberation (thanks to an anonymous referee for raising this issue). Rationalism typically describes making up one's mind as a 'here and now' type of process (see also Jongepier (2017)). To follow a suggestion made by Jonathan Lear, it may be worth considering the notion of what he calls "extended practical deliberation" (Lear 2004, 453).

[19] Though see Boyle's (2015) reply to the worry regarding triviality.

[20] A normative rationalist doesn't have to say transparency is all that matters, in Patty Hearst's case or others. I discuss this reply and in particular its implications (namely, if we bring in other values, transparency may hardly ever be what really matters) below.

[21] In Strijbos and Jongepier (2018), we argue that for instance making up one's mind in a negative or depressed 'state of mind' makes a dysfunctional outlook on the world more salient, and can thus make it harder for an individual to consider healthier or less self-destructive outlooks on the world. The latter, importantly, are attitudes and outlooks the individual also and would realize she has if only she would look *at* her attitudes a little more, rather than look transparently through them.

[22] For an insightful article on the various conceptions of authenticity and implied conceptions of alienation, see Feldman and Hazlet (2013).

10). So why not just think these are different yet compatible conceptions of alienation? Patty Hearst might be alienated in one sense (hypothetical resistance), but not in another (conforms to transparency).

I have two responses to this worry. First, even if there are distinct conceptions of alienation at play here that are compatible, it might still be the case that one of those types of non-alienation is more valuable to have than the other. One might still prefer, for normative reasons, a non-transparency-based conception of alienation. For example, would it not be better for Patty to find out whether or not she would (counterfactually) have feelings of resistance, rather than finding out whether she has a genuinely transparent (or secretly opaque) outlook on the world? Isn't the former type of self-knowledge more worth having in her case? To be sure, the rationalist doesn't have to say that transparent self-knowledge is always valuable, no matter what.[23] Also, an inferentialist can easily concede that making up one's mind can be valuable in some cases and not in others. Indeed, some sort of pluralism about which types of self-knowledge are valuable seems to be the right methodology here. The worry is that the relative value of transparent self-knowledge in our actual lives might turn out to be at best modest and at worst negligible. If there are other types of alienation to choose from, this might make the situation worse for rationalists, not better.

Second, it's not clear whether the two (or more) conceptions can actually be made compatible at all. There's reason to think that the rationalist's claims about transparent self-knowledge are precisely meant to compete with rival conceptions of alienation. In an article on Harry Frankfurt's work, Moran (2002) presents his own transparency approach as the one to be preferred, rather than *complementing* Frankfurt's view. Also, recently both Alec Hinshelwood (2013) and Michael Garnett (2013) have plausibly interpreted Moran's account as offering a competing theory of authenticity and/or autonomy. So the false positive objection does not necessarily confuse two types of alienation so much as force us to think about whether the two can be made compatible and if not, which one is to be preferred. If nothing else, the false positive objection thus puts pressure on the normative rationalist to make explicit why non-alienated self-knowledge understood as transparent self-knowledge is worth having, compared to other varieties of non-alienated self-knowledge.

Another possible reply a rationalist might give to the false positive objection is to say that the objection simply misfires because Patty did not really make up her mind in the first place. Not being able to make up one's mind is just part of the definition of being brainwashed. And if Patty did not really make up her mind, she would still be alienated even on rationalist terms, hence there would not be any false positives.

But one wonders: what might deliberating be if not weighing reasons pro and con and coming to a judgment, which Patty Hearst can do? Moreover, it's important not to forget that rationalists define alienation in terms of having an empirical, theoretical self-perspective on oneself or one's mental states. The only way for the rationalist to say that Patty was alienated would be by saying she in fact adopted a theoretical, third-person perspective on herself (despite appearances to the contrary). This is hard to accept, if only because it would make the notions of 'making up one's mind' and adopting a 'third-personal stance' much more technical and revisionary than rationalist will, I reckon, be prepared to accept. Also, it is worth pointing out that Patty Hearst's case appears to generalize rather easily to other cases, such as deliberation as it occurs in circumstances involving manipulation, coercion, and oppression. In those circumstances, it is much harder – and potentially patronizing – to say that these

[23] Thanks to an anonymous referee for bringing this up.

individuals are not capable of making up their minds either. And yet these circumstances, too, can be a recipe for alienation and self-deception (see e.g. Mackenzie 2002).[24]

And yet there seems to be something right about the thought that Patty Hearst and persons in similar circumstances have not really made up their minds at all. How do we account for this intuition? I believe this intuition has to do with the fact that such individuals are not deliberating 'well' or are not deliberating with what we might say are 'reasons of their own'. In other words, what we worry about when we worry about Patty Hearst is not *whether* she deliberated or made up her mind, but rather *how* she did. Indeed, it seems to me that saying something along these lines provides the most promising way for the normative rationalist to deal with hard cases.

However, notice that this does mean that one will have to focus on what we might call the 'quality of deliberation', which is not without costs for the rationalist. If rationalists take this route, then acquiring transparent self-knowledge no longer simply requires world-directed deliberation, as is normally suggested. It now turns out that it requires deliberation of a specific type, or deliberation in specific circumstances, or being able to deliberate with 'good' reasons – whatever that turns out to mean. Going down this route thus requires rationalists to give up on their proceduralist ambition to remain neutral as to what counts as the relevant 'input' to the deliberative procedure. It requires making much more than Moran's preferred "modest claims" about there being "logical room" for raising deliberative questions or a mere "power" to make up one's mind.

Let's suppose that rationalists are willing to abandon their proceduralist methodology, in order to save normative rationalism. Then what might such an alternative, non-procedural version of rationalism look like? To answer this question, I propose, in the next section, to consider some interesting parallels with the debate on personal autonomy. In that debate, which is concerned with the question of what it means to govern oneself, feminist theorists have criticized a dominant 'proceduralist' approach and have in its place developed a 'relational' alternative. Perhaps there's a relational way out for normative rationalism.

5 Relational Rationalism

A so-called proceduralist account of autonomy defines autonomy in terms of certain reflective capacities of an individual.[25] Influential proceduralist theories include Frankfurt's account, according to which one is autonomous if and only if one is capable of high-order identification with one's first-order desires (Frankfurt 1971; see Dworkin 1988a for a somewhat similar view). Proceduralist theories of autonomy are sometimes also referred to as a 'content-neutral' accounts, because they do not pose any substantive constraints on the content of a person's

[24] How exactly we should understand the relationship between e.g. oppression and having (non-alienated) self-knowledge is a mighty difficult question. I certainly don't want to claim that being oppressed means one lacks non-alienated self-knowledge (Khader 2012). The point here merely is that these circumstances could create epistemic obstacles for non-alienated self-knowledge, and rationalism in its current form wouldn't be able to explain how.

[25] Much has been written on the topic of autonomy and how (not) to define it. For the purposes of this paper, however, it is not necessary to go into all of the interpretations and distinctions, given that the aim here is primarily methodological: to see whether the debate concerning transparent self-knowledge can benefit from considering some of the moves made in the debate on autonomy. For a good overview, see (J. S. Taylor 2005).

mental states (such as involving lack of self-respect) nor on the circumstances in which the relevant capacity is exercised.

Proceduralist theorists disagree about what the relevant reflective capacity must be, but they typically agree that having and exercising the relevant capacity is necessary and sufficient for autonomy (Mackenzie and Stoljar 2000, 14). As Natalie Stoljar writes, on procedural accounts in general, "there is no reason in principle why choosing subservience, or adopting oppressive norms, could not be autonomous" (Stoljar 2015). Dworkin for instance consider a person who "wants to conduct his or her life in accordance with the following: Do whatever my mother or my buddies or my leader or my priest tells me to do". Such a person, Dworkin claims, "counts, in my view, as autonomous" (Dworkin 1988b, 21).

A recurring objection to proceduralist views raised by feminist/relational theorists is that someone's capacity for rational reflection can be hijacked by various unjust or oppressive socio-political influences (Mackenzie 2002). The feminist literature on autonomy is rife with examples purporting to show that reflective identification or reflection cannot be sufficient for autonomy.[26] As Marina Oshana puts it, on a relational but not a proceduralist account of autonomy, "it is possible for two individuals to satisfy all the psychological and historical conditions (...) but to differ with respect to their status as autonomous beings—and this difference is to be explained in terms of some variance in their social circumstances" (1998). Whereas proceduralist theories focus mainly on the details of the relevant rational capacities, relational theorists instead focus on the various social, cultural and political conditions that need to be in place in order to develop such capacities at all and what is required in order to sustain them.

It should be clear that the proceduralist's position looks a lot like the normative rationalist's position. In fact, the false positive objection I introduced earlier can be understood as an instantiation of the sort of objections or problem cases that have been directed against proceduralist theories of autonomy. So the thought here is this: if it makes sense to construe autonomy along relational lines, in order to answer false positive-type objections, it might make sense to try to construe a relational version of rationalism.

The basic relational rationalist thesis, as I imagine it, will come down to something like this: for the transparency method to be conducive of non-alienated self-knowledge, it has to be followed in the 'right' circumstances. The 'bad' circumstances might involve bad reasoning (problems with the process) or bad deliberative material (problems with the input to the reasoning process). The false positive objection would not apply to relational rationalism, because making up one's mind is not meant to be sufficient for acquiring non-alienated self-knowledge. On a relational rationalist view, when we are told that a subject has "made up her mind", we are not thereby told that she has acquired non-alienated self-knowledge. What we furthermore need to know is how, where, and why she made up her mind, and probably a lot more.

In the next section I will discuss a number of objections to relational rationalism, but here I want to briefly respond to what I regard as an illegitimate concern with relational rationalism, which will also help to clarify the relational rationalist project. The worry is that developing a relational rationalist account would be ad hoc. One might wonder: aren't we now just tweaking normative rationalism *such that* it can handle hard cases? Talk of the 'right' deliberative

[26] Influential examples include the Deferential Wife discussed by Thomas Hill (1991), the Taliban Woman discussed by Marina Oshana (Oshana 1998), and Mackenzie's discussion of 'Felicity Porcelline'.

circumstances was never part of Moran's original account, nor indeed of much of the literature that followed.

Though this may be true, 'going relational' isn't an ad hoc move. This has to do with the fact that the notion of 'rationality' is a normative notion. Just as beings can be a-moral (they lack the relevant moral capacities) as a well as immoral (they have the moral capacities but exercise them badly), beings can be a-rational as well as irrational. If we apply this to rationalism, we get the following:

(1) *a-rational*: someone does not follow the transparency procedure
(2) *irrational:* someone follows the transparency procedure, but does so 'badly' or 'unsuccessfully'

Most defences of rationalism so far have focused on a version of (1). Teleological rationalism is concerned with our having a capacity for transparent question-settling, in contrast to beings that lack this capacity, such as chairs and cats and survivors of a frontal lobotomy. Moran's own claims about alienation focus on a different version of (1), namely, on subjects who do not 'obey transparency' (at all) on *specific occasions*, such as the person undergoing psychotherapy who inferred her mental states (though not a survivor of a frontal lobotomy). By contrast, a relational rationalist would be interested in exploring (2): beings who clearly have the capacity for obeying transparency, and who actually do obey transparency on various or all occasions, but who do so 'badly' in some way. Not much has been said about option (2) in the literature.

Having put the ad hoc worry aside, the key question for the relational rationalist is: how must we understand its appeal to good-making circumstances of deliberation? The circumstances in which one (fails to) acquire self-knowledge are hardly ever the central topic of philosophical discussion. Still, some philosophers have made careful moves in this direction. William Alston for instance mentions that if a self-ascription is made "in a fit of abstraction, its indicative value will be impaired if not altogether lost" (Alston 1965). Johannes Roessler adds that the epistemic value is also undermined if made "during sleep, or in a state of advanced intoxication" (Roessler 2015). And Victoria McGeer writes that self-ascriptions are authoritative "assuming I am sane, and sincere, and not deeply distracted" (McGeer 2007, 81). Thus, some philosophers at least recognize that if one is asleep, drunk, deeply distracted or in a psychosis, then following some method of self-knowledge probably will not actually lead to self-knowledge.

A relational rationalist would say that most of us are – outside of the philosophy room, anyway – familiar with various bad-making circumstances of self-knowledge. Having a grumpy outlook on the world, for instance after waking up in a bad mood, might lead one to have a misleading view of what one wants. The same goes for being in pain, high or low on hormones, hungry, grieving, in love, excited, annoyed, high on caffeine, low on caffeine, merely tipsy, and so on. They all matter to how you make up your mind and what the epistemic value of your self-ascription is.

But how should we incorporate these insights into a theory of non-alienated self-knowledge? Such a view is currently missing. A relational rationalist might follow Rawls' strategy with respect to the method of reflective equilibrium, and appeal to "considered judgments":

> considered judgments (...) enter as those judgments [which] are most likely to be displayed without distortion (...) For example, we can discard those judgments made with hesitation, or in which we have little confidence. Similarly, those given when we

> are upset or frightened, or when we stand to gain one way or the other can be left aside. All these judgments are likely to be erroneous (...). (Rawls 1999, 42)

In a Rawlsian spirit, the relational rationalist might say that only one's "considered reasons" are permissible material for entering the transparency procedure. But the relational rationalist will have to go one step further and exclude not just the reasons that occur in hesitant, upset, or fearful moments, but also those that are the result of brainwashing, and possibly also those arising in contexts of oppression, coercion and manipulation.[27] And then of course there's the hormones and the caffeine.

This is just a very rough sketch of the shape relational rationalism might take. However, we can already note two important advantages of such a view. First, it is much better able to deal with the false positive objection compared to procedural rationalism. It has a much richer conceptual apparatus to deal with hard cases. Second, relational rationalism is a substantially normative account: having non-alienated self-knowledge in the relational rationalist sense really does appear to be worth having and striving towards. However, relational rationalism faces two new – and rather big – problems.

6 Trouble for Relational Rationalism

The first problem for relational rationalism is that 'rationalism' and 'relationalism' actually turn out to form an unhappy marriage. The rationalist's intellectual passion is directed at questions like 'what sort of rational activity *is* making up one's mind exactly?' or, 'how should the relevant world-directed questions be characterized when it comes to mental states other than belief?' (Boyle 2015, 340). But it is unlikely that *these* questions will inspire the relational theorist. The relationalist will instead be interested in questions such as: 'what are the circumstances in which making up one's mind – whatever that involves, exactly– actually provide one with non-alienated self-knowledge?' or, 'can socio-political circumstances also be bad-making and if so, when exactly?'

Relationalists and rationalists thus have very different ideas about where the philosophical action should be. A relationalist will no doubt want to try to find out which conditions are epistemically undermining, which ones are epistemically beneficial, and which ones make no difference to non-alienated self-knowledge either way. For relational theorists, 'transparency' might not be a *wrong* answer to the question of what non-alienated self-knowledge requires, just as 'yeast' is not a *wrong* answer to the question of what one needs in order to make bread. But it is, from the relationalist's perspective, a disappointingly incomplete answer. To put the point bluntly: if we're interested in intriguing cases of self-knowledge, then it seems we don't have much reason to be particularly interested in rationalism.

Relational rationalism faces another problem, which I'll call the 'particularist problem'. As we've seen, the relational rationalist will have to appeal to good-making circumstances of non-alienated self-knowledge, for instance by appealing to Rawlsian considered reasons. But the problem is that there appears to be no principled, non-question begging way of defining what the good-making circumstances would be. Let me explain.

[27] Clearly this move of extension would not be available to Rawls because the method of reflective equilibrium is meant to be a method of moral justification so this would beg the question. But given that relational rationalists have different goal, this strategy might be available to them.

A natural answer to the question of what the right circumstances are in which making up one's mind is conducive of non-alienated self-knowledge is to say: if you're self-deceived, then making up your mind will not provide you with non-alienated self-knowledge. But as Annalisa Coliva rightly notes, if we were to add "self-deception" to our list of epistemically undermining conditions, as Crispin Wright (1989) for instance has done, then such a theory would not "have much of a point" (Coliva 2009, 372).[28] We need more than a condition that comes down to defining what the 'good-making circumstances' are in terms of the absence of non-alienation.

The only other strategy is to try to come up with a list of what we normally consider to be 'suspicious' circumstances, as Rawls does. But this strategy is not very promising either. This is because suspicious moods or circumstances, though perhaps they are often epistemically undermining, *need* not be (cf. de Maagt 2016). Being tipsy or drunk might also lead to self-insight (as the saying goes: 'in vino veritas'). Being angry, too, can allow one to suddenly — though authentically and transparently – realize what one wants or believes in. Consider an oppressed housewife who's suddenly had enough, and angrily "looks out onto the world" and comes to new and authentic insights about her mental states precisely because of – not in spite of – her angry outlook on the world.

We sometimes have good reason to trust, rather than distrust, the reasons that occur to us in our less reflective or calm states of mind, such as when we are overwhelmed by emotions – *especially* as far as self-knowledge is concerned. Some have even argued that emotions hold the key to acquiring (true) self-knowledge (Mackenzie 2002). Even if one thinks this claim is too strong, our less calm and emotional outlooks on the world can be perfectly legitimate, and there's nothing epistemically untrustworthy about them in principle. The same holds true of the other supposedly suspicious circumstances that one might put on the list, such as grief, sadness, being hormonal, in love, intoxicated, and so on. All of those circumstances can bring great insight into one's (non-alienated) mental states. It also is not clear whether being manipulated or oppressed necessarily leads to lesser self-knowledge and greater alienation. All of this very much depends on the *particulars* of the case (hence I call it the particularist worry). Any list, therefore, that a relational rationalist will come up with, will fail to offer principles that are true in general.

All of this puts the normative rationalist in an uncomfortable position. If what the rationalist was after was to be able to deal with standard objections by "going normative", then relational rationalism is to be preferred over its teleological and proceduralist cousins. But in spite of the strengths of relational rationalism, it faces the challenge of articulating the good-making circumstances of transparency in a non-question begging way, as well as avoiding the scenario of rationalism itself dropping out of the equation. Perhaps it's time to reconsider whether going for normative immunity is really the most promising way of dealing with standard objections.

Concluding Remarks: What's Next?

The primary aim of this paper was to distinguish between different possible normative conceptions of rationalism. First, I reconstructed what I referred to as teleological rationalism, according to which having the capacity to make up one's mind is part of our essential rational nature and necessary for living a worthwhile life. Next, I discussed procedural rationalism and

[28] John Christman similarly claims that the type of self-reflection necessary for autonomy must involve "no self-deception" (1991, 11).

relational rationalism, both of which are accounts of how making up one's mind is related to a distinctly valuable type of self-knowledge, namely, non-alienated self-knowledge.

These different conceptions of normative rationalism have not been distinguished in the literature, not even by rationalists themselves. Moran, for instance, moves from claiming that he meant to say "something the denial of which would be equivalent to denying that people ever actually reason to a conclusion" (2004, 458) to claiming there is something "*wrong* with [a] person's access to himself" when "he cannot consciously avow the first attitude and can only ascribe it to himself on the evidence" (2001, 86 emphasis added). I've argued these are in fact distinct claims, and must be evaluated separately. Doing so constituted the second aim of this paper.

Having laid out and criticized all three versions of normative rationalism, one might wonder: isn't normative rationalism simply engaged in a different kind of project compared to other theories of self-knowledge?[29] One could indeed say that normative rationalism on the one hand and theories such as inferentialism on the other are strictly speaking not competitors. That may in itself be a valuable lesson to learn, and could even make possible a type pluralism in the self-knowledge debate that is currently missing. But notice that this requires thinking, first, that rationalism isn't also an epistemic theory of self-knowledge (which it clearly is), and second, that non-rationalist theories of self-knowledge couldn't compete with rationalism on a normative level.

It might make a lot more sense to insist that all these theories *are* engaged in the same project and to proceed to examine precisely the normative significance of the other players. What prudential or moral value does, say, self-interpretation or self-expression have? How does normative rationalism fare relative to *other* normative theories of self-knowledge? As Ryle wrote, a diary is "a valuable source of information about the diarist's character, wits and career" (Ryle 1949, 149). Rereading one's diary – or one's social media timeline or self-tracking app observations or what have you – might give one valuable insight into what one (really) wants, believes or intends. In fact, it's not inconceivable that not just transparency but also *opacity* has "a genuine role to play in accounting for aspects of the structure and phenomenology of moral experience" (Moran 2001, 193). We can hold people accountable if they make up their minds too often and self-objectify too little. In short, normative rationalism is a genuine competitor in the larger game of normatively significant theories of self-knowledge, and there is value in asking how well it fares, relative to the alternatives. It's not just that epistemically-minded critics of rationalism have failed to take seriously the rationalist's distinctly normative claims, they have also failed to take consider the normative potential of their own theories.

In any case, my discussion of the three different conceptions of normative rationalism has, regrettably, been mostly negative. At the same time, it should be clear that none of the objections I have raised present a knock-down argument against any version of normative rationalism. Rather, the concerns are best understood as an attempt to clarify which steps must be taken for normative rationalism to succeed. So, where should the normative rationalist go from here?

First, rationalists need to make explicit how normatively ambitious rationalism is meant to be and which normative claims they are (not) making. Do normative rationalists want to make both teleological claims about our essential rational nature *and* more substantially normative claims about the value of transparent self-knowledge?

[29] Thanks to an anonymous referee for pressing me on this point.

Second, if normative rationalists want to make both claims, then they must say more about the relation between the two. Ideally, the normative rationalist would provide an account of how one would get from the claim that there is something 'wrong', in a teleological or transcendental sense, with (human) beings who lack the power to make up their minds, and the claim that there is something 'wrong', in a more substantial sense, with Quantified Self-advocates, inferentialists about self-knowledge, or persons who infer they're angry or sad in the process of psychotherapy or conversation. The slide from claims about rational essences to moral norms and values of transparency and opacity must be accounted for.

Third, normative rationalists need to explicitly engage with the wider literature on autonomy, authenticity, alienation, and bad faith. Also, the relative value of making up one's mind versus treating oneself non-transparently as a thinker who just happens to have certain mental states, needs to be addressed. Only then would rationalists and critics of rationalism be in a position to distinguish between the different ways in which various types of self-knowledge might be valuable, and assess their relative normative strengths and weaknesses.

A final observation I want to make is that relationalism may still have a lot going for it, regardless of rationalism. The first problem with relational rationalism after all was due to its connection with rationalism, and the second problem can arguably be overcome by giving up the search for general good-making circumstances of (non-alienated) self-knowledge and instead embrace the epistemically capricious nature of self-knowledge. Even though the particularist problem as I called it might be a problem for rationalism, it can also be taken as an opportunity to begin to explore a new, particularist approach to self-knowledge. I cannot develop the particularist approach to self-knowledge here,[30] but let me end by giving a rough indication of where one might start.

Particularist theories of moral reasons and moral responsibility (Dancy 2004; Vargas 2013) are typically contrasted with so-called atomist theories. An atomist about reasons for instance holds that "any feature that is a reason in favour of action in one case will always be a reason in favour of action wherever it occurs" (Dancy 2004, 772). On a particularist view, by contrast, it "all depends on the circumstances" (ibid.). In a similar vein, an atomist (or what I have called proceduralist) about self-knowledge would claim that whether or not some method (be it transparent question-settling, self-expression, or self-interpretation) is knowledge-conducive depends on whether the relevant method was followed, and does not depend in any interesting way on the (bodily, psychological, social) circumstances in which the method was followed. Or, as Manuel Vargas puts it, on particularist view (and I am here replacing 'free will' with 'self-knowledge'), "we cannot answer the question of whether an agent has [self-knowledge] simply by looking at the agent. What we need to know are facts about both the agent and the circumstances." (Vargas 2013, 206).

Perhaps what we should say, then, is this: the question of which circumstances are the 'right' circumstances or methods for acquiring valuable types self-knowledge, *just depends*. It depends on the person one is and the life one leads. It depends on one's mood, on where one is, who one is talking to (if anyone at all), the stakes involved, what one has done and what one will do, how one treats oneself, what one knows, it depends on one's dreams and fears and much else besides.

Philosophers typically don't think too highly of 'it depends'-answers, given that such answers are theoretically somewhat unsatisfying. And perhaps they are. But outside of academia, especially as far as self-knowledge is concerned, it really all does just depend,

[30] See Jongepier (2017).

and in real life it makes little sense to think of (non-alienated) self-knowledge as something one necessarily acquires through method Q or in circumstances XYZ.

In the end, it may be difficult for a normative rationalist, or any theorist of self-knowledge really, to say anything more substantial than "it depends" without begging relevant questions. That may be bad news for the self-knowledge debate, but I doubt it is bad news – indeed probably no news at all – for most self-knowing and self-deceived agents.

Acknowledgements Work on this paper was supported by the Niels Stensen Foundation. I would like to thank Quassim Cassam, Naomi Kloosterboer, Lukas Schwengerer, Robin Wisse and participants of the Cambridge Mind Seminar for helpful comments and discussion. Most of all I want to thank Sem de Maagt.

References

Alston WP (1965) Expressing. In *Philosophy in America: Essays*, edited by max black and William Payne Alston. Ithaca, NY: Cornell University Press

Bagnoli C (2007) The Authority of Reflection. Theoria 58:43–52

Borgoni C (forthcoming) Basic self-knowledge and transparency. Synthese 1–18

Boyle M (2011) 'Making up your mind' and the activity of reason. Philosopher's Imprint 11(17):1–24

Boyle M (2012) Essentially rational animals. In rethinking epistemology, ed. Guenther Abel and James Conant. Walter de Grutyer, Berlin

Boyle M (2015) Critical study: Cassam on self-knowledge for humans. Eur J Philos 23(2):337–348. https://doi.org/10.1111/ejop.12117

Byrne A (2018) Transparency and self-knowledge. Oxford University Press

Carman T (2003) First persons: on Richard Moran's authority and estrangement. Inquiry 46(3):395–408. https://doi.org/10.1080/00201740310002424

Cassam Q (1987) Transcendental arguments, transcendental synthesis and transcendental idealism. Philos Q 37(149):355–378

Cassam Q (2014) Self-knowledge for humans. Oxford University Press, Oxford

Cassam Q (2015) What asymmetry? Knowledge of self, knowledge of others, and the Inferentialist challenge. Synthese 194:1–19. https://doi.org/10.1007/s11229-015-0772-7

Christman J (1991) Autonomy and personal history. Can J Philos 21(1):1–24

Christman J (2005) Autonomy, self-knowledge, and Liberal legitimacy. In *Autonomy and the Challenges to Liberalism* Cambridge: Cambridge University Press https://doi.org/10.1017/CBO9780511610325.016

Coliva A (2009) Self-knowledge and commitments. Synthese 171(3):365–375

Dancy J (2004) Ethics without principles. Clarendon Press, New York

Driver J (2001) Uneasy virtue. Cambridge University Press, Cambridge

Dworkin G (1988a) The theory and practice of autonomy. Cambridge University Press, Cambridge

Dworkin G (1988b) The theory and practice of autonomy. Cambridge University Press

Evans G (1982) The varieties of reference. Clarendon Press, Oxford

Feldman SD, Hazlett A (2013) Authenticity and self-knowledge. Dialectica 67(2):157–181. https://doi.org/10.1111/1746-8361.12022

Finkelstein DH (2012) From transparency to Expressivism. In *Rethinking Epistemology*, edited by Günter Abel and James Conant. Berlin: Walter de Gruyter

Frankfurt HG (1971) Freedom of the will and the concept of a person. J Philos 68(1):5–20. https://doi.org/10.2307/2024717

Garnett M (2013) Taking the self out of self-rule. Ethical Theory Moral Pract 16(1):21–33

Gertler B (2011) Self-knowledge. Routledge, New York
Golob S (2015) Self-Knowledge, Agency and Self-Authorship. Proceedings of the 15 Aristotelian Society XCV (3)
Haase M, Mayr E (2019) Varieties of Constitutivism. Philos Explor 22(2):95–97. https://doi.org/10.1080/13869795.2019.1601754
Hill TE (1991) Autonomy and self-respect. Cambridge University Press, Cambridge
Hinshelwood A (2013) The relations between agency, identification, and alienation. Philos Explor 16(3):243–258. https://doi.org/10.1080/13869795.2013.815260
Hunter E (1956) Brainwashing: the story of men who defied it. World Distributors, London
Jones K (2005) Moral epistemology. In *The Oxford Handbook of Contemporary Philosophy*, edited by frank Jackson and Michael smith. Oxford: Oxford University Press
Jongepier F (2017) The circumstances of self-knowledge. PhD dissertation, Radboud University Nijmegen, Nijmegen
Khader SJ (2012) Must Theorising about adaptive preferences deny Women's agency? J Appl Philos 29(4):302–317. https://doi.org/10.1111/j.1468-5930.2012.00575.x
Kloosterboer N (2015) Transparent emotions? A critical analysis of Moran's transparency claim. Philos Explor 18(2):246–258
Korsgaard CM (2008) The constitution of agency: essays on practical reason and moral psychology. The Constitution of Agency. Oxford University Press, Oxford
Lear J (2004) Avowal and Unfreedom. Philos Phenomenol Res 69(2):448–454
de Maagt S (2016) Reflective equilibrium and moral objectivity. Inquiry 60(5):1–27. https://doi.org/10.1080/0020174X.2016.1175377
Mackenzie C (2002) Critical reflection, self-knowledge, and the emotions. Philos Explor 5(3):186–206. https://doi.org/10.1080/10002002108538732
Mackenzie C, Stoljar N (2000) Relational autonomy: feminist perspectives on autonomy, agency, and the social self. Oxford University Press
McGeer V (1996) Is 'self-knowledge' an empirical problem? Renegotiating the space of philosophical explanation. *Journal of Philosophy* XCIII 10:483–515
McGeer V (2007) The moral development of first-person authority. Eur J Philos 16(1):81–108. https://doi.org/10.1111/j.1468-0378.2007.00266.x
Moran R (2001) Authority and estrangement: an essay on self-knowledge. Princeton University Press, Princeton
Moran R (2002) Contours of agency: essays on themes from Harry Frankfurt. In *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah buss and lee Overton. Cambridge: MIT Press
Moran R (2004) Replies to heal, Reginster, Wilson, and Lear. Philos Phenomenol Res 69(2):455–472
Nussbaum M (1995) Aristotle on human nature and the foundations of ethics. World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams
Oshana M (1998) Personal autonomy and society. J Soc Philos 29(1):81–102
Rawls, John. 1999. A theory of justice, Revised Edition. Harvard University Press
Roessler J (2015) Self-knowledge and communication. Philos Explor 17 (2)
Ryle G (1949) The concept of mind. Routledge, London
Shah N, Velleman JD (2005) Doxastic deliberation. The Philosophical Review 114(4):497–534
Shoemaker S (2003) Moran on self-knowledge. Eur J Philos 11(3):391–401. https://doi.org/10.1111/1468-0378.00192
Stoljar, Natalie. 2015. "Feminist perspectives on autonomy." In *The Stanford Encyclopedia of Philosophy*
Strijbos D, Jongepier F (2018) Self-knowledge in psychotherapy: adopting a dual perspective on one's own mental states. Philosophy, Psychiatry and Psychology
Taylor JS (2005) Personal autonomy: new essays on personal autonomy and its role in contemporary moral philosophy. Cambridge University Press, Cambridge
Taylor K (2006) Brainwashing: the science of thought control. Oxford University Press, Oxford
Tugendhat E (1986) Self-consciousness and self-determination. MIT Press, Cambridge
Vargas M (2013) Building better beings: a theory of moral responsibility. Oxford University Press, Oxford
Wright C (1989) Wittgenstein's rule-following considerations and the central project of theoretical linguistics. In *Reflections on Chomsky*, ed. George Blackwell