Attitude and Commitment

Kloosterboer, N.J.G.

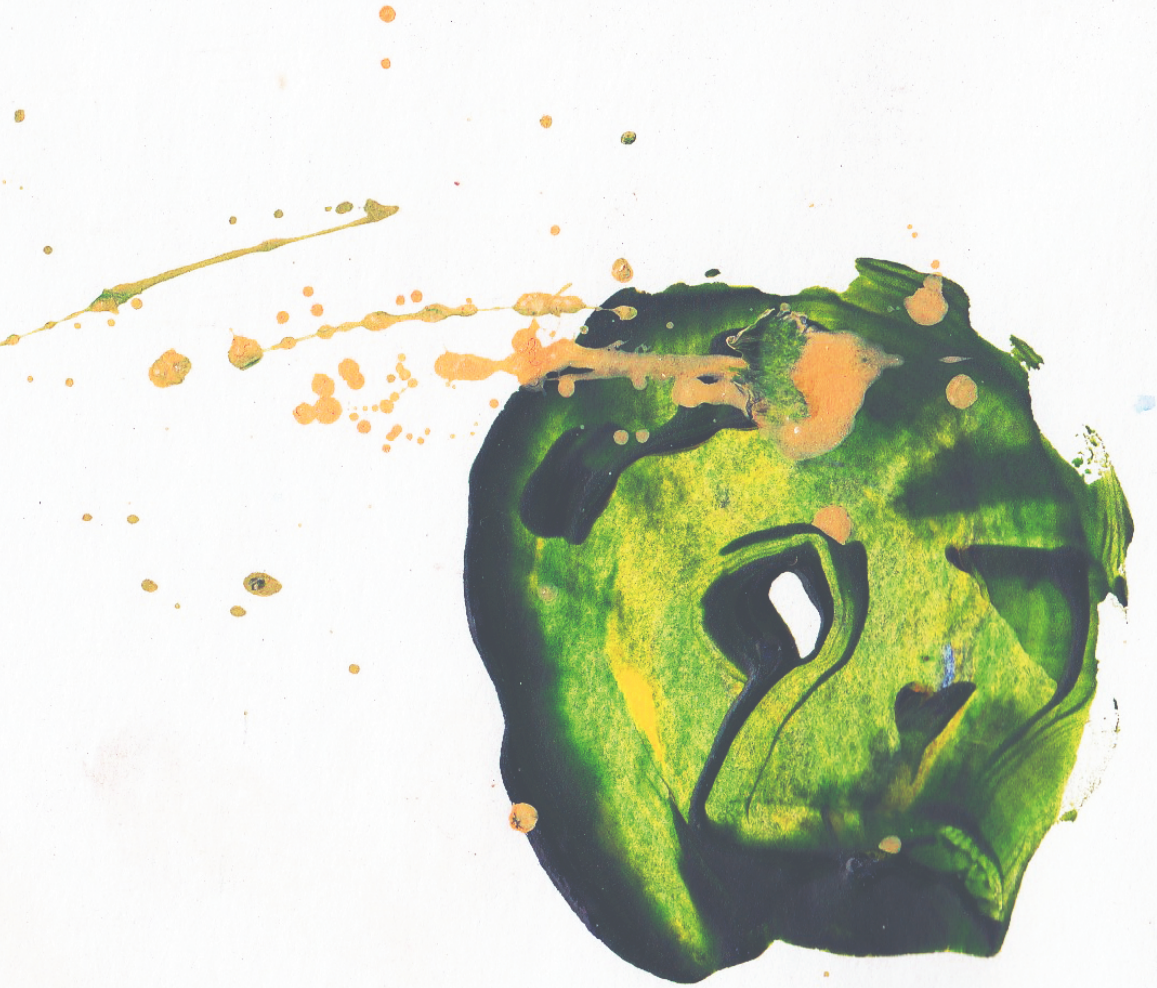
2019

document version
Publisher's PDF, also known as Version of record

Naomi Kloosterboer

ATTITUDE AND COMMITMENT

a study of transparent self-knowledge



ATTITUDE AND COMMITMENT

A Study of Transparent Self-Knowledge

Naomi Kloosterboer

Cover Design: Marieke de Wit
Cover Image: Naomi Kloosterboer
Printed by Ipskamp B.V.

VRIJE UNIVERSITEIT

ATTITUDE AND COMMITMENT

A Study of Transparent Self-Knowledge

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geesteswetenschappen
op vrijdag 25 oktober 2019 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Naomi Jacoba Gerarda Kloosterboer

geboren te Amsterdam

promotoren:	prof.dr. R. van Woudenberg
	prof.dr. G. Glas
copromotor:	dr.ing. L.C. de Bruin

CONTENTS

DANKWOORD

Het is geweldig dat ik zo lang heb kunnen studeren en de stimulans heb gekregen om dit proefschrift te schrijven. Maar zonder ontzettend veel hulp van anderen was het er niet geweest. Vooral de voltooiing voelde voor mij als acrobatiek.

Eigenlijk wilde ik dit dankwoord met een geschikt citaat beginnen, zoals jij, René, een aantal keer voor de inleiding gesuggereerd hebt. Bijvoorbeeld een citaat dat de lezer prikkelt; dat de lezer meeneemt in het idee van transparante zelfkennis; een poëtische noot bij de droge academische kost. Maar verder *no pressure*… Terwijl druk op mij leggen natuurlijk het laatste is wat jij zou willen! Al vanaf het begin straal je vertrouwen uit. En plezier. Ik heb ontzettend veel van alle feedback, meetings en conferentiebezoeken geleerd. Je filosofische precisie en secure gebruik van concepten hebben enorm bijgedragen aan mijn zorgvuldigheid in het filosofisch schrijven en argumenteren. Onze persoonlijk band en de tact waarmee je precies op het juiste moment een mailtje stuurt met 'Hoe gaat het?' zijn voor mij erg belangrijk geweest. Ik had me geen fijnere promotor kunnen wensen. Gerrit, dank voor de interessante gesprekken en de feedback die je hebt gegeven. Ik zie ernaar uit je boek te lezen en onze discussies voort te zetten. Leon, je bent een meester in de pragmatische aanpak en zo'n aanpak kwam vaak als geroepen. Dank voor alle betrokkenheid.

Aan alle collega's en vrienden op het filosofie-departement van de Vrije Universiteit Amsterdam, dank voor het stimulerende klimaat, de lunches, waar iedere dag iets nieuws te leren valt over *powerfood* én het ouderschap, maar vooral de ontspannen sfeer. Voor feedback op mijn werk en de paper-meetings die uit een uitzonderlijke combinatie van precisie en (filosofische) jeu bestonden, wil ik in het bijzonder Emanuel, Geertjan, Hans, Jeroen, Josephine, Lieven, Rik, Scott, Tamarinde, Thirza, Valentin en Wout bedanken. Jan Willem, ik heb ontzettend veel geleerd van ons duo-paper en kijk uit naar een vervolg. Door het delen van PhD- en baby-sores (en de nodige biertjes), heb ik me ontzettend gesteund gevoeld door Irma, Jojanneke, Judith, Marije, Phil, en de rest van de VUFIL app.

Sinds februari werk ik op de Universiteit Utrecht en ik ben dankbaar voor de behulpzame sfeer waarin ik terecht gekomen ben. In het bijzonder bedank ik

Annemarie, Eric, Jesse, Jos, Marcus, Mariëtte, Mathijs, Naomi, Niels, Pepijn, Sander en Sem voor alle hulp en hartelijkheid. Andere collega's in Nederland en Vlaanderen die ik graag wil bedanken voor hun bijdrage aan mijn werk en plezier in filosofie zijn Fleur Jongepier, Dawa Ometto, Sabine Roesser, Katrien Schaubroeck, Maureen Sie, de PhD's betrokken bij de OZSW en vele anderen. Beate, je hebt een belangrijke rol gespeeld in mijn vertrouwen om een PhD te ambiëren. Dank voor alle betrokkenheid en feedback op mijn werk. Pieter, jij weet als geen ander hoe je studenten kunt betrekken bij een discussie. Dank voor de steun tijdens mijn studie.

In 2015, I have had the opportunity to spend some time at the University of Birmingham. Thanks are due to Lisa Bortolotti for making this possible, for directing project PERFECT and the stimulating reading groups, and for helpful feedback. I felt very welcome at the department. I am also grateful to Matt Boyle and the philosophy department at Harvard University for making my research visit in 2016 possible. Thank you, Matt, for the inspiring discussions and encouraging feedback. Dick Moran, thank you for writing *Authority and Estrangement*, which has been a crucial motivation for this dissertation. I also very much enjoyed the classes on philosophy of action, and the lunch where we discussed my initial thoughts for the last chapter. Thank you, Cassirer family and Elena, for your warmth during my stay. Other people from abroad that have helped me along the way are Quassim Cassam, Kelvin McQueen, Johannes Roessler, David Widerker, and especially Ryan Cox for diligent and helpful feedback.

Lieve vrienden, zonder jullie was ik allang knettergek of een kluizenaar. Het proefschrift slokte weliswaar bijna al mijn tijd op, maar als ik jullie zag, maakte dat niets meer uit. Nonnen en kootz, voetbal en ouwehoeren blijft de beste manier om filosofie op pauze te zetten. Voor je lettermagie, ook op deze kaft, wil ik Marieke bedanken. Louise, hoe de dag op de VU ook was; koffie met jou maakte 'm sowieso beter. Yara, dichtbij of (relatief) ver, ik ben zo blij dat ik je altijd kan vinden.

Lieve paranimfen: Lieke, alles wat bij de VU-collega's staat geldt ook voor jou, maar dan in het kwadraat. Ik ben ontzettend blij dat je, als mijn *partner in crime*, straks naast me staat. Jolien, ik vind het nog steeds heel leuk dat onze vriendschap begonnen is tijdens het bespreken van onze gedeelde interesse in Harry Frankfurt. Als een deel van onze vrijheid gelegen ligt in de houding die we aannemen, dan kijk ik die van jou graag af. Zo ook tijdens de promotie!

Lieve familie, (schoon)broers en zussen, ik ben heel dankbaar voor het mooie *patchwork* dat wij vormen. Oma Gerda, opa Jan en oma Coby, ik vind het een feest dat ik dit met jullie kan delen. Sally, wat heb ik geluk met een tante die, als editor en

taalfanaat, altijd bereid is om, humor incluis (poor froggy…), mijn stukken te corrigeren.

Lieve ouders, cadeau-ouders en schoonouders, ik ervaar ontzettend veel steun van jullie. De zorg voor Mischa en het plezier dat daarvan afstraalt is hartverwarmend. Pap, als er iemand de bron is voor mijn interesse in autonomie en *agency*, dan moet jij dat wel zijn. Mam, je hebt me geleerd wat het is om veerkrachtig te zijn en door te zetten. Lieve Leen, wat had ik deze mijlpaal graag met je gedeeld.

Dan als laatste mijn twee liefste mannen. Mischa, het is heerlijk om me in jouw wereld te wanen en zo al het werk te vergeten. Fabian, dank dat je met mij een team wilt zijn en dat je zo anders bent dan ik.

INTRODUCTION

Self-Knowledge, Science, and Agency

1. Introduction

Do I believe that it is raining? Do I believe that my partner and I will grow old together? Do I intend to pay back the money I borrowed? Will I attend my friend's birthday party? Do I like strawberries more than raspberries? Do I want another child? Do I value family over work? Should I focus on having fun, being a parent, a career woman, a good friend? These kinds of questions, both the more trivial and the more substantial ones, are central to this dissertation. They are the kinds of questions whose answers, if true, provide one with a piece of self-knowledge, namely, self-knowledge of one's own *intentional mental attitudes*. How do we answer such questions? Why would such answers count as self-knowledge? And why are these questions important for us? Are they related to the idea that we are agents – i.e., persons with a sense of self-direction, responsibility and commitment?

These questions inform the basis of this dissertation and belong to the philosophy of self-knowledge. A principal point of departure in thinking about self-knowledge is the difference between knowledge of one's own mental attitudes and the mental attitudes of others. This difference is predominantly viewed in light of its epistemological features: a person seems to have so-called *privileged* and *peculiar access* to her own mental attitudes but not to those of others (cf. Byrne 2005; Gertler 2015). Privileged, because self-knowledge seems especially epistemically secure, and peculiar, because it is only the person herself who has this kind of access to her own mental attitudes. Self-knowledge is thus often viewed as epistemologically distinct in light of its epistemic status, namely particularly secure, and in light of the method used to acquire self-knowledge, namely a method only available to the person whose mental life is at issue.

What is often left out of these discussions of the epistemology of self-knowledge is the connection between self-knowledge and the nature of the person who is seeking self-knowledge. Why should self-knowledge matter to us? What are the connections of self-knowledge to personhood, to moral psychology, and to (mental) agency? These moral psychological issues about self-knowledge aren't meant to replace questions about its epistemology, but they do provide a renewed starting point to ask and answer them. The underlying thought is that moral psychological issues determine the kind of self-knowledge at issue, and that, in turn, the kind of self-knowledge at issue influences the relevant answers to the epistemological issues.

This is the principal context for the questions pursued in this dissertation. The dissertation consists of a collection of papers that inquire how self-knowledge should be understood within a moral psychological framework, where the connections to personhood, moral psychology, and (mental) agency are recognized as crucial to understanding the nature of self-knowledge. Especially, this dissertation seeks to incorporate into the picture of self-knowledge the connection between having a mental attitude and being committed to the view of the world inherent in the attitude. I will dub the kind of self-knowledge that "respects" this connection between attitude and commitment *transparent self-knowledge*. A central thought that will be developed in this dissertation is that committing oneself and sticking to one's commitments is to manifest one's *agency*. It requires taking up a certain responsibility for who one is and what one stands for. This responsibility establishes a certain kind of freedom: it is up to us what we stand for. At the same time, it may also be portrayed as a burden. Since our mental attitudes express our commitments, we cannot escape this responsibility – we *must*[1] take a stance, even if we would sometimes like to avoid it. Following this line of thought, the main idea of the dissertation is that *acquiring transparent self-knowledge involves manifesting one's agency*.

1.1 Science versus everyday life

My study of transparent self-knowledge has been partly born out of amazement at the increasingly all-encompassing scientific perspective on the nature of the human mind. The achievements and developments of the behavioral, cognitive and

[1] The sense of "cannot" and "must" that is used here is to be understood in the following way: given our abilities of committing ourselves and taking a stance, the options available to us have changed in such a way that each option we choose is informed by this ability. The sense used here might thus be called "agentive" (instead of say, purely metaphysical or purely normative) and follows the spirit of recent work on the agentive modalities. cf. Maier (2013). More on these abilities further on in this introduction.

neurosciences are so extensive that 'science has come...to be widely viewed as the primary source of concepts and theories sufficient to describe and explain all of reality including human beings' (Haldane 2012, 672). The scientific view of human beings no longer seems to be one view among many but has become absolute, and it seems to make the viewpoint of lived human experience obsolete. Science is often regarded as the sole authority to tell us about the nature of things. This faith in science and disregard for other forms of knowledge is part of a larger historical movement called *scientism*.[2]

What is, roughly, the view of self-knowledge in science? In science, our capacity for self-knowledge is heavily doubted and sometimes even declared an illusion.[3] To give an idea of this position, consider the following characterization of our capacity for self-knowledge by Daniel Dennett:

> …each of us is in most regards a sort of inveterate auto-psychologist, effortlessly *inventing* intentional interpretations of our own actions in an inseparable mix of confabulation, retrospective self-justification, and (on occasion, no doubt) good theorizing. (Dennett 1987, 91; emphasis in original)

Self-attributions of mental attitudes, in this view, are nothing more than the result of theorizing about what could go on in our minds that would explain what we do. Moreover, the suggestion is that being right about what goes on in our minds is, for a large part, a matter of luck and not tied to any privileged or peculiar position of the first-person.

There is an enormous contrast between this science-driven view of self-knowledge and the intuitions and practices regarding self-knowledge in everyday life. If the science-driven view is correct, we should be very skeptical about our own and each other's claims to self-knowledge. In everyday life, however, we don't seem to be skeptical at all. Perhaps we doubt our own and each other's claims about our deepest desires or about the significance of traumatic experiences, but self-knowledge isn't limited to such claims about our deeper psychological make-up. Self-knowledge also concerns very mundane beliefs, desires and intentions. By and

---

[2] cf. Schöttler (2012). As van Woudenberg et al. (2018) write: "Scientism is, roughly, the view that only science can provide us with knowledge or rational belief, that only science can tell us what exists, and that only science can effectively address our moral and existential questions." For proponents of scientism, see Dennett (2017); Ladyman (2011); Rosenberg (2011). For an excellent overview of different kinds of scientism, see Peels (2018). For different kinds of explorations and criticisms of scientism, see de Ridder, Peels, and van Woudenberg (2018).

[3] cf. Curruthers (2011); Dennett (1987); Wilson (2002); Nisbett & Wilson (1977); Lamme (2011).

large, we trust our own and each other's claims to such ordinary self-knowledge. If you were to tell me, in the relevant circumstances, that you intended to go to bed early, that you would like to go to the new Wes Anderson movie, or that you are walking to the shop to buy basil for tonight's supper, I would readily believe you. What's more, we use one another's self-ascriptions in our plans and actions. For instance, if you tell me that you intend to meet me tomorrow at noon at a nice coffee place, I plan to be there and expect to meet you, thereby using your self-ascription as information as to where you will actually be tomorrow at noon. Contrary to the science-driven view, then, in everyday life we seem to take for granted that we have self-knowledge.

Given scientism, it is clear how to solve this conflict between science and *common sense* on whether to trust our everyday claims to self-knowledge. That is, since scientism holds that only science can provide us with knowledge, we should discount our everyday intuitions. The authority of science to settle the matter also derives support from history. After all, science has shown us that our solar system is heliocentric, that biological species evolve, that brain damage can cause personality change, etcetera, thereby undermining everyday intuitions. But should the scientific view of self-knowledge prevail over our everyday intuitions and practices regarding self-knowledge? That doesn't follow from the scientific claims alone. For scientific results can only be interpreted relative to claims as to what self-knowledge is. Such claims, however, aren't scientific but philosophical in nature. This means that we have to engage in philosophical work to determine what view of self-knowledge we *should* use in evaluating the scientific results.

Another reason to not let science uncritically trump our common sense intuitions is that the view of self-knowledge purported by science bypasses precisely those things about self-knowledge that seem to be its most distinguishing features, namely the relation between self-knowledge, agency, and the first-person perspective. In science, the prevalent view of self-knowledge is the *observational model*. Both in the history of philosophy and from a scientific perspective, this has been the dominant model to understand self-knowledge. It is a model based on the idea that introspection is a form of observation: it is to look – or observe – what is inside one's mind.[4]

There are two fundamental problems with the observational model. The first problem is that, in the observational model, the difference between a person's

[4] There are different accounts of self-knowledge elaborating this basic idea, varying from so-called inner sense accounts, to monitoring accounts and to functional accounts. Each of these accounts has its particular merits and difficulties in explaining self-knowledge. For the most prominent accounts, see Armstrong (1968); Lycan (1996); Goldman (1993; 2006); Nichols and Stich (2003); Rosenthal (2005).

knowledge of her own and of other's mental attitudes is merely a matter of epistemic access. Only the person herself has observational access to her own mental attitudes. But this is just a contingent fact about our current capacities. We can easily envisage the difference dissolving. For instance, we might develop telepathic powers, thereby acquiring the ability to observe other minds. Or perhaps we develop a scanning mechanism that not only scans our own minds but also the minds of our interlocutors (e.g., by wiring our brains). Observational access thus doesn't bear out an essential difference between self-knowledge and knowledge of someone else's mental attitudes. As Richard Moran writes:

> One thing that is unsatisfying about any perceptual [i.e., observational] model of self-consciousness is that perception is a relation that, in principle, should be possible with respect to a whole range of phenomena of a certain type. On such a model, then, there would seem to be no deep reason why one couldn't bear this quasi-perceptual relation to the mental life of another person as well as oneself. (Moran 2001, 33)

Why do we need such a "deep reason" that distinguishes between self-knowledge and knowledge of someone else's mental attitudes? Without such a deep or essential difference, it is a genuine (future) possibility that someone else will tell you what you believe, want, etcetera. This seems irreconcilable with the idea that a person's attitudes express her view of things and that she has certain reflective capacities such as making up her mind and taking a stance. Hence, on the assumption that there should be a deep difference between knowledge of one's own and of other mental attitudes, explicating the difference in terms of observational access will not suffice (cf. Shoemaker 1996; Moran 2001).

The second problem is that, in the observational model, the relation between having a mental attitude and having knowledge of it is merely causal and contingent. An implication of this is a kind of splintered view of our mental capacities. In this view, the fact that we have mental attitudes, that we can reason, make plans, wonder about what is important in life and so forth, and the fact that we have a capacity to know of our mental attitudes and that we are self-conscious, are in principle independent of one another. If these capacities are in principle independent, then it should be possible that a person has no knowledge of any of her mental attitudes, without this lack of knowledge making any difference to the rest of her mental life. However, such *self-blindness* appears to be a conceptual impossibility (cf. Shoemaker 1996, Lecture II). Being self-blind involves being able to have and conceive certain

mental states, but not being able to have peculiar access to them, i.e., to have knowledge of them through a first-person method. To see why self-blindness appears to be conceptually impossible, consider the following examples. Suppose that we are talking about what to cook for dinner tonight and you say "We should make *melanzane alla parmigiana*!" But when I subsequently ask you whether you like "melanzane", you say you don't have a clue. This would be quite absurd, for why would you then make the exclamation about "melanzane" at all? Or consider a person walking to the medicine cabinet and taking pain killers and responding, when asked whether she is in pain, that she doesn't know, but given what she is doing, she probably is. Both cases seem inconceivable, because saying that we should make "melanzane" and taking painkillers doesn't show us what mental attitudes a person has, rather they are intelligible only if she knows that she likes "melanzane" or knows that she is in pain.[5] It thus seems incomprehensible for a person to be able to have certain mental states, such as being in pain, *and* not being able to know of them first-personally. If such self-blindness is inconceivable in this way, this means that the conception of ourselves as agents *entails* a capacity for self-awareness: individuals cannot make sense of the idea of being an agent without any form of self-apprehension.[6]

The outlook of this dissertation is to neither uncritically side with the scientific doubt about self-knowledge nor with the dominant observational model of self-knowledge. Rather, I will be attempting to do justice to our everyday intuitions surrounding self-knowledge and to understand philosophically how the nature of the first-person agential perspective should inform our conception of self-knowledge. Peter Carruthers (2011, xiii) claims that 'it is a mistake to address questions in the philosophy of mind in [a non-naturalistic way]… [and] even more misguided to address them in ignorance of the relevant data in cognitive science, as many philosophers continue to do.' I agree. But I also think it is a mistake to regard the scientific results as clear-cut evidence that speaks for itself. This dissertation aims to show that self-knowledge is the kind of human capacity that merits close philosophical attention and that the philosophical analysis should drive the interpretation of the scientific results.

[5] For a complete exposition of the example of pain, see Shoemaker (1996, 273-5).

[6] There is, of course, more to say about the observational model, about its merits as well as about its problems. In this dissertation, however, I follow these arguments and assume that the observational model cannot explain our intuitions regarding self-knowledge.

1.2 Agential account of self-knowledge

Recent developments in the self-knowledge debate seem to offer a promising alternative to the observational model and the scientific perspective. These developments show a restoration of the importance of the first-person perspective for understanding self-knowledge. This is seen, especially, in so-called agential accounts of self-knowledge (cf. Bilgrami 2006; McGeer 1996; Moran 2001; Tugendhat 1986). One prominent agential account of self-knowledge was developed by Richard Moran (2001) in his influential monograph *Authority and Estrangement: An Essay on Self-knowledge*. Moran's account is a crucial point of departure for this dissertation. To understand the questions pursued, it is therefore helpful to give a short outline of Moran's account.

Moran explicitly wants to move the discussion of self-knowledge 'from the epistemology of introspection to a set of issues in the moral psychology of the first-person' (2001, 4). Self-knowledge, according to Moran, is a person's knowledge of her own mental attitudes, but more than that it is her knowledge of her mental goings-on *from her own perspective*, that is, her first-person perspective. A person isn't merely witnessing her mental life; the mental goings-on under investigation are her *own* beliefs, intentions, emotions, desires, etcetera. It should matter to her what they are. It should matter to her, because such attitudes express her view on things, i.e., what she takes to be true, what she will do, how she feels and what she wants. '[W]hatever "self-knowledge" of the relevant kind is,' as Moran (2001, 136-7) writes, 'it should be something we can understand as having a special *importance* to the person, an importance beyond the usefulness of having some way of knowing, for example, one's own parentage or tax bracket.'

Moran's account of self-knowledge seeks to accommodate a number of asymmetries between how we know our own minds and how we know the minds of others. Knowing one's own mind seems to have an authority, directness and non-evidential basis that one's knowledge of other minds lacks.[7] Moran seeks to account for these asymmetries by developing the idea that a person's relation to her mental life is different from her relation to the mental life of others (or of other's relation to her mental life). This essential difference is, according to Moran, not a difference in privileged access but a difference in the way a person is involved in her own mental life, namely as mental agent.

[7] That self-knowledge *seems* to have these features is generally acknowledged and many philosophers consider them as a datum a theory of self-knowledge should explain, which minimally implies that the theory should explain why self-knowledge appears to have these features, while knowledge of other minds doesn't.

Moran seeks to capture the special relation of the subject to her own mental attitudes by distinguishing between two different stances or perspectives one can take towards one's own mental life: a "theoretical" or "third-person" stance and a "deliberative" or "first-person" stance. These stances correspond with two kinds of questions about and inquiries into one's mental life. A theoretical question about one's mental life is 'one that is answered by discovery of the fact of which one was ignorant', Moran explains, 'whereas a practical or deliberative question is answered by a decision or commitment of some sort, and it is not a response to ignorance of some antecedent fact about oneself' (2001, 58). The core idea of the *deliberative stance* is that my relation to my mental attitudes is not that of an expert witness or a bystander who happens to have the best information about my mental attitudes. Different from being a witness or bystander, I do not merely register what is present in my mind. Rather, my mental attitudes express my relation to the world, my *stance* or *grasp* of how things stand in the world.[8] As such, they must be seen by me 'as expressive of [my] various and evolving relations to [my] environment, and not as a mere succession of representations (to which, for some reason, [I am] the only witness)' (Moran 2001, 32).[9]

An essential feature of this picture of self-knowledge and the deliberative stance is that self-knowledge is, for Moran, not a matter of arriving at the most accurate description of my psychological state, but a matter of *avowal*. An avowal consists of a report of one's mental attitude including an explicit endorsement of its content. In the case of belief, to avow my belief is to express my 'own present commitment to the truth of the proposition in question' (Moran 2001, 86). By avowing myself on the matter, I take responsibility for my mental attitude. Avowing the belief that *p* thus expresses my endorsement of *p*. Moreover, it involves a commitment to the truth of *p*. As soon as I start doubting *p*'s truth or as soon as I

[8] Importantly, this claim holds for self-knowledge of *intentional mental attitudes*. These attitudes, such as beliefs, emotions, desires and intentions, are fundamentally different from sensations, headaches and heart rates, because they involve, for the subject of those states, a characteristic grasp of the world. That is to say that these attitudes involve, from a first-person perspective, grasping the (propositional) object of those states *as true*, *as to be done*, *as dangerous*, etcetera.

[9] The deliberative stance is partly motivated by Moore's paradox. Although it is not unusual to say about someone else that she believes something that is actually false, and although it could well be a state one is actually in, from a first-person point of view, it does not make sense to say "P but I don't believe P"(see Moore 1993; Moran 2001, Ch. 3). Why is that? As Moran seeks to explain, the best explanation for the paradoxality of Moorean sentences is that, from the first-person perspective, there is a certain blindness to the difference between declaring one's belief that p and declaring p itself. As Wittgenstein (2009 [1953]) remarks, 'if there were a verb meaning "to believe falsely," it would not have any significant first-person present-tense indicative' (quoted in cf. Moran 2001, 73). As Moran (2001, 77) puts it: 'What is unavoidable from the first-person perspective… is the connection between the question about some psychological matter of fact and a commitment to something that goes beyond the psychological facts.'

reconsider the issue, the avowal ceases to exist (Moran 2001, 74-7, 80-2). Taking this responsibility is what Moran envisages as taking a deliberative stance toward our mental life.

A person arrives at an avowal of her mental attitudes, according to Moran, by answering a relevant question about the content of the attitude. A person is, so to say, drawn to the content of the attitude, because her attitudes are transparent: in a way to be specified, she looks beyond (or "through") the attitude to what the attitude is about. This is why Moran's account of self-knowledge is called a "transparency" account. As Moran writes, the basic idea of transparency is that '[w]hen asked "Do I believe P?", I can answer this question by consideration of the reasons in favor of P itself' (Moran 2003, 405; see also Moran 2001, 62-3). When a person is asked, for instance, whether she believes that it is raining, she will, rather than looking for evidence about having the belief, look out into the world to see whether there are signs of rain (or revisit the weather forecast she heard on the radio, etcetera). Based on her considerations about the weather, she makes up her mind and avows the belief that it is raining. This is how she comes to know her mind.

This is a brief overview of the basic tenets of Moran's account, which is a paradigm exemplar of agential accounts of self-knowledge in general. Numerous questions arise at this point. These are questions about the relation between avowal and self-attribution, about the connection between making up one's mind and self-knowledge, about the nature of the deliberative stance, the mental agency involved, the scope of the account, and the kind of self-knowledge that agential accounts seek to depict. Given the threat from science to our everyday intuitions about self-knowledge, how we answer these questions adjudicates our capacity for self-knowledge. In what follows, I will introduce these questions by staging five themes that are central to this dissertation. These themes should be read as providing, not a systematic overview of the chapters, but a common basis underlying them. After the discussion of the themes, I will provide an overview of the chapters to come.

2. Transparency

As mentioned, the proposed agential view of self-knowledge is indebted to the idea of transparency. In philosophy of mind and its history, the notion of "transparency" is used with different connotations. I will discuss three of these connotations, so as

to introduce the topic and set the stage for the discussion of (agential) transparency accounts of self-knowledge.[10]

One sense of transparency is the Cartesian one: it is to use the notion of transparency to depict the relation between the subject and her mind. The mind is, so to say, transparent to itself. Transparency here denotes the self-intimating nature of all mental items in Descartes' view of the mind, where self-intimacy means that 'if the mind is in a certain state, the subject necessarily knows that she is in that state' (Paul 2014, 295).[11] Since Freud's discovery of the unconscious and the more recent research on unconscious influences on our minds, this rather strong thesis of the Cartesian transparent mind has had few (or it would seem even no) champions. This notion of transparency as self-intimacy isn't used in this dissertation. Still, the idea of privileged access, a view widely shared in the self-knowledge debate, might be seen as a weaker version of self-intimacy (cf. Carruthers 2011; Paul 2014). Although it doesn't include that any mental state a person is in is known by her, it does mean that a person herself is in a good position, and especially a position better than others, to know her own mind.

Another sense of transparency is what is also called the *diaphanousness* of experience and concerns a phenomenological thesis about experience (cf. Stoljar 2004). An experience is transparent in this sense (and thus diaphanous) if, even if we try to pay closer attention to the experience itself, we only seem able to attend to what we experience and not that we experience it (cf. Dretske 2003). Similarly, we might say that a mental attitude is transparent in this sense if, even if we try to pay closer attention to the attitude itself, we only seem to be able to attend to what the attitude is about and not that we hold the attitude – to its attitudinal quality, so to say. This use of the term comes from G.E. Moore's paper "The Refutation of Idealism," where he writes the following about the perception of the blue sky:

> ... that which makes the sensation of blue a mental fact seems to escape us: it seems, if I may use a metaphor, to be transparent – we look through it and see nothing but the blue ... (Moore 1903, 446).

[10] I largely follow Sarah Paul's (2014) terrific overview in "The Transparency of Mind."

[11] Gertler (2015) defines self-intimacy (or omniscience) as requiring that 'being in a mental state suffices for knowing one is in that state.' But self-intimacy seems to be a stronger thesis than that: it places a necessary connection between being in a mental state and knowing this is so. There is no mental state one is in but of which one is ignorant. Self-intimacy should be distinguished from the infallibility thesis, which says that if a subject believes she is in a mental state, she is in that state. Or, alternatively put, the infallibility thesis holds that 'one cannot have a false belief to the effect that one is in a certain mental state' (Gertler 2015).

Phenomenologically, it thus seems that we cannot, as such, attend to our experience, only to the object of our experience. This phenomenological thesis can be related to a metaphysical thesis as well as to an epistemological thesis. The metaphysical thesis associated with the diaphanousness of experience is that experience, and mental attitudes as well, don't have introspectable properties. The associated epistemological thesis is that there is no direct information about our minds available to us: for if we cannot attend to the mental sensation of the blue, but only to the blue sky, then what information are we supposed to employ in order to know that we experience the blue sky? On this reading, the diaphanousness of experience (and mental attitudes) turns self-knowledge into a puzzle.

This latter reading is closely related to the last sense of transparency that I want to discuss, which is the one central to this dissertation and which will be meant in using the word "transparency." Where diaphanousness refers to the phenomenological or metaphysical character of awareness itself, the latter notion of transparency refers to how a person achieves self-knowledge. It is associated with Wittgenstein's criticism of traditional views of introspection and the relation of a person to her own mind that such views support. Put differently, it is associated with a renewed conceptualization of the first-person perspective. As a notion pertaining to what we do when we self-ascribe or answer questions about mental attitudes (and, in some cases, also experiences), it is only natural that it inspired the current transparency accounts in the self-knowledge literature, including Moran's account. The most famous expression of this notion of transparency can be found in Gareth Evans' work:

> ... in 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?' I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p ... (Evans 1982, 225).

What Evans describes here is 'how a question about one's own belief must present itself, from the first-person point of view' (Moran 2001, 60). And in general terms, we might say that what is transparent in these quotations is one's own thinking: when asked a question about one's own state of mind, one's "gaze", i.e., one's thinking and attention, isn't directed "inward", to the mental state itself, but

"outward", to the intentional object of one's mental state. On the other hand, we could also say that, from a first-person perspective, the question about my own mental attitude is transparent to a question about its intentional object, for example, whether I believe that *p* is transparent to the question whether *p*. These characterizations of this notion of transparency are two sides of the same coin: in the former you start from the mind-directed question and end up characterizing the direction of thought as transparent, i.e., at the object of one's thinking, and in the latter you start from the first-person perspective, i.e., the thinking mind, and end up characterizing the relation between the two questions as transparent.

We get into muddy waters, however, when we try to state this notion of transparency in less general terms. It is especially unclear whether we should understand it as making an empirical claim, i.e., that our thinking will be directed "outward," as a normative claim, i.e., that for conceptual, epistemic or practical reasons our thinking must be directed "outward," or yet as another kind of claim related to our agential capacities. Hence, one issue with transparency accounts is how to formulate a canonical formulation of it.

3. Two Topics Problem

Another problem comes into view if we consider the epistemic basis of transparent self-knowledge. If self-knowledge is to be *knowledge,* then it should have an epistemic basis. But what is the epistemic basis in transparency accounts of self-knowledge? Formulating such epistemic credentials in transparency accounts is particularly difficult because they face the so-called *puzzle of transparency* or the *Two Topics Problem*. For what is supposed to link the question about the worldly state of affairs to the question about one's own mental attitude? How can a person intelligibly answer the question about her mental attitude by answering a question about its content? Or, put in different terms, how can she use the content of an attitude to self-attribute that attitude?

On the face of it, the intentional object of a mental attitude doesn't provide information about the attitude itself. For instance, the fact that it is raining doesn't entail that I believe that it is raining, nor provides evidence for it: one can imagine numerous scenarios in which it is raining but I do not believe that it is, or in which I believe that it is raining but it isn't. In other words, the proposition that it is raining and the belief that it is raining neither stand in a relation of implication nor of evidential support. This is why transparency accounts face the *Two Topics Problem* (TTP): the problem that the apparent basis for self-knowledge, i.e. *p* (including

evidence in favor of *p*), doesn't provide a reason to self-ascribe a particular mental attitude regarding *p*.[12]

Traditionally, an epistemic basis is supposed to be obtained either through observation or through reasoning.[13] As outlined above, my discussion of self-knowledge assumes that the kind of self-knowledge that we are after isn't the result of observation. Moreover, it embraces the idea that self-knowledge is, or at least appears to be, direct. If self-knowledge actually *were* direct, it wouldn't be the result of reasoning. But if self-knowledge only *appears* to be direct, then it might be the result of reasoning. Would an appeal to reasoning enable us to solve TTP and secure an epistemic basis for transparent self-knowledge?

A prominent transparency account that claims that reasoning would indeed solve TTP is Alex Byrne's inferential account of transparent self-knowledge. Byrne (2005, 2011, 2018) holds that a person acquires self-knowledge by making an inference from "world to mind," e.g., in the case of belief, an inference from *p* to *I believe that* p. As just explained, prima facie, this seems a very odd thing to do, for the world contains little information about a person's mind. But Byrne, although admitting that such an inference is neither deductively valid nor inductively strong, maintains that such an inference is knowledge-conducive. The reason for this is that, as Byrne claims, the inference is self-verifying: moving from *p* to I believe that *p* implies the truth of the latter because, says Byrne, 'inference from a premise entails belief in that premise' (2011, 206). Byrne thus needs this assumption in order to make the inference from *p* to *I believe that p* epistemically justified.

But should we accept this assumption? This question ultimately rests on another question. Namely, what should be true of reasoning to verify this assumption? Byrne presupposes that when a person reasons, she believes the premises. This is in line with current orthodox views in the philosophy of reasoning, in which reasoning is analyzed as a change in attitudes.[14] In this view, reasoning indeed necessarily involves believing the premises and conclusion, and so this view supports Byrne's assumption. The question is, however, whether this view is true. If it isn't, then Byrne's assumption, which is crucial in his account, would be undermined. Assessing Byrne's account thus involves assessing the view of

[12] The problem at hand is also known as the puzzle of transparency, the problem of two subject matters, and the evidentialist objection. See, for instance, Barnett (2015), Byrne (2005), Gallois (1996), Martin (1998), Moran (2001), O'Brien (2003), Roessler (2013a), among others. Since its recognition, a number of transparency accounts are proposed specifically in response to this problem.

[13] This refers to a puzzle created by Boghossian, who states that knowledge is the result of (1) observation, of (2) reasoning, or of (3) nothing. If (1) and (2) were both rejected, it seems that a deflationary account of self-knowledge is all that is left. cf. Boghossian (1989).

[14] Cf. Boghossian (2014); Broome (2013); Harman (1986); McHugh & Way (2018).

reasoning as "change in attitudes." The questions whether self-knowledge can be the result of reasoning will thus be answered in this dissertation by inquiring the nature of reasoning.

4. Attitudes and scope

The third theme in this dissertation is the scope of Moran's, or an agential style, transparency account of self-knowledge. Moran explicates his transparency account for beliefs, but claims that the account, with relevant adjustments, should apply to mental attitudes other than belief as well.[15] There have been a lot of questions, however, about the precise scope of Moran's account.[16] Does it apply to all kinds of beliefs? Does it apply to mental attitudes such as emotion, desire and care? Does it apply only to conscious attitudes and rational attitudes, or across the board? In this section, I want to draw some basic distinctions that appear in different chapters to come and provide a framework for assessing the scope of Moran's, or an agential style, account.

4.1 Intentional mental attitudes

First of all, Moran's account applies only to intentional mental attitudes, such as beliefs, intentions, desires, many kinds of emotions, etcetera. It isn't an account of self-knowledge of sensations, character traits, one's identity nor one's personal history. The reason for this is that a person is only actively involved in the way explicated by Moran's account in the case of intentional mental attitudes. As Moran writes:

> A distinguishing fact about "intentionally characterized" phenomena generally (not only states of mind, but actions, practices and institutions, including linguistic ones) is that they admit of a distinction between inside and outside perspectives, the conception of them from the point of view of agents or participants as contrasted with the various possible descriptions in some more purely naturalistic or extensional idiom. (Moran 2001, 34-5)

[15] See, for instance, Moran (2001, 64-5; 2004, 471; 2012, 214).
[16] As can be witnessed in many responses to Moran on this topic. See Ashwell (2013a); Cassam (2011); Gertler (2011); Lawlor (2009); Shah & Velleman (2005); and many more.

The agential "inside" perspective that agential transparency accounts seek to capture is manifest only in intentional mental attitudes. Sensations also allow for an "inside" perspective, but here the inside perspective is qualitative rather than agential in nature. A person cannot, for instance, feel pain for a reason or against reason. This, among other things, means that the subject, as an agent, presumably remains passive with respect to her sensations.[17]

4.2 Doxastic, conative and affective

Intentional mental attitudes come in different varieties. Importantly, there are differences in *kinds* of attitudes: there are *doxastic* attitudes, such as belief, *conative* attitudes, such as intentions and desires, and *affective* attitudes, such as emotions. As already noted, Moran claims that his account applies across the board. But what works for doxastic attitudes need not work for conative or affective attitudes, and vice versa. It is important, then, to investigate the possibility of extrapolating an account of self-knowledge of belief (as most current transparency accounts are) to other kinds of mental attitudes.

4.3 Types of attitudes

There are also different types of attitudes.[18] A first distinction that should be made is between *occurrent* and *standing* attitudes. Suppose I believe that I am taller than six feet. Sometimes this belief might actually be expressed in my thoughts, and as such may "occur" to me, but more often it will not figure in my thoughts at all. Still, it would be silly to say that if it isn't on my mind, then I don't believe it. Rather, we might say it is a standing belief: a belief that one has that isn't playing any role in one's mind at the moment. Other attitudes, too, can be occurrent or standing. Sally can desire to take a long vacation in two months, even if that desire isn't manifest all the time. If Sally has the desire, but it isn't manifest, it is a standing desire. If it is manifest, it is occurrent. The case of emotion is more complex. Some emotions might only exist in an occurrent form, e.g., rage, while others might exist as a standing attitude. For instance, you can be angry at someone for, say, a week, while also going to work and having fun with colleagues. Perhaps, when you think about the person you are angry with, your anger again becomes manifest. One can thus be angry even if the anger doesn't occupy one's mind all the time. In such a case, it is a standing attitude.[19]

---

[17] For more on the difference between self-knowledge of intentional mental attitudes and sensations, see Boyle (2009b).

[18] These distinctions are mainly based on cf. Schroeder (2015); Schwitzgebel (2015).

[19] Some philosophers use the adjectives "occurrent" and "standing" to distinguish between different kind

Next, we should distinguish between *explicit* and *tacit* attitudes and between *explicit* and *implicit* attitudes. The belief that I am taller than six feet is explicit, and might be occurrent or standing, if I have formed a belief with that exact content. But if I believe that I am taller than six feet, it seems natural to say that I also believe that I am taller than five feet, taller than four feet, that I have a height, etcetera. But these contents needn't have ever crossed my mind. Such beliefs are called *tacit beliefs* (also known as *dispositions to believe*). Even if I have never explicitly formed a belief with these precise contents, I should be in a position to easily, and perhaps automatically, derive their contents from my explicit belief that I am taller than six feet. Talk of tacit attitudes is most common in the case of belief, but I don't see why we shouldn't apply the notion to other attitudes as well.

Another contrast with an attitude being explicit is that it is implicit. *Implicit attitudes* often conflict with explicit attitudes and as such are often thought to be revealed by emotional reactions rather than by reflection or introspection. An example of such an implicit attitude is that a person, despite her explicit non-racist commitments, behaves less rigidly and feels safer with, e.g., Caucasians than with non-Caucasians. There is a lot of discussion about the nature and implications of implicit attitudes. Some doubt whether such implicit states should be identified as attitudes, others doubt whether there is anything unified enough to be called a state, yet others use the concept of implicit attitudes to undermine the possibility of genuine self-knowledge. For now, though, this is just meant as a brief introduction to the distinction.

Finally, there is the distinction between *conscious* and *unconscious* attitudes. There are two ways of understanding this distinction, one having to do with what is currently attended to and the other with what can be attended to, as contrasted with sub-personal processes or states that cannot be the focus of attention in the same way. On a first reading, calling an attitude conscious is calling it an explicit occurrent[20] conscious attitude, i.e., an attitude that one is currently aware of. In this

---

of mental items. Occurrent mental states, they claim, are things such as sensations, thoughts, and mental acts (e.g. judging). By contrast, standing mental states, i.e. the intentional mental attitudes discussed so far, have a dispositional nature. In this picture of attitudes, a person holds a certain attitude only if she expresses the attitude in a particular range of actions and reactions (cf. Cassam 2014; Schwitzgebel 2010). The strange result of this position is that there is no such thing, for instance, as a "conscious belief." A belief could be expressed in a conscious judgment (an occurrent mental state) but could not become conscious itself (cf. Boyle 2009a). This seems to me a counterintuitive consequence and one that we should seek to avoid. This is not to say that I dismiss the relevance of patterns of actions and reactions to self-knowledge of our intentional mental attitudes. But I do think that what that relevance precisely is, should be explicated in different terms.

[20] Occurrent is not the same as conscious: one might be in pain (an occurrent mental state) but being so focused on something else (perhaps winning a race) that one is conscious only of the finish line.

reading, an unconscious attitude is an attitude one isn't currently aware of. On a second reading, one that was instigated by the "discovery of the unconscious" by Freud, a conscious mental state is one that can be brought to one's attention (i.e., the explicit and tacit standing attitudes) and an unconscious mental state is one that cannot be brought to one's attention, at least not without an enormous amount of therapeutic work (a mental state or mental attitude that resides in the Freudian unconsciousness). Agential transparency accounts of self-knowledge seek to explain how conscious – in the Freudian sense – mental attitudes can become conscious, i.e. the focus of attention. That is, how explicit and tacit standing attitudes can become occurrent attitudes one is aware of.

4.4 Rationality of attitudes

Another distinction to be discussed is the rationality of mental attitudes. Let me first outline in which ways an attitude may lack rationality. Brute likes and dislikes, for instance, lack rationality because they are a-rational, which means that the question of being justified by reasons doesn't apply to them. Secondly, an attitude might lack rationality if a subject's apparent justification for holding the attitude is false. In such a case of an irrational attitude, the subject takes herself to have reasons in favor of the attitude that either aren't factual or actually don't speak in favor of the attitude (in this case, from the subject's point of view, her attitude *is* rational). Another possibility is that an attitude lacks rationality if the subject maintains holding the attitude despite her best judgment to the contrary (this need not imply she doesn't have *any* reason in favor of holding the attitude). Such attitudes are called recalcitrant or persistent. A fourth option for lacking rationality is an alienated attitude. An attitude is alienated if the subject isn't in touch with reasons pertaining to the attitude; if her reflection on the object of the attitude doesn't seem to be related to having the attitude. In other words, we can distinguish between the following categories: attitudes that lack rationality can be *a-rational*, *irrational*, *recalcitrant*, or *alienated*.

In each of these cases, the counterpart attitude can be called rational. Hence, by rational attitude we might mean an attitude: (i) whose justification depends on reasons (not a-rational); (ii) that is actually justified by good reasons (not irrational); (iii) that isn't contrary to the subject's reasons (not recalcitrant); or (iv) whose reasons the subject is in touch with (not alienated).

For transparent self-knowledge, this latter notion of rationality is the most important one. What matters for transparency is whether a person is *in touch with* reasons that she takes to be pertinent to the mental attitude in question. If a person, for instance, comes to know of her unconscious resentment through psychotherapy,

she might still lack transparent self-knowledge if she cannot avow, and thereby endorse the view purported by her resentment.[21] The notion of alienation thus relates to the connection between self-attributing a mental attitude and being committed to the view purported by the attitude.

4.5 Trivial vs. substantial

The last distinction that I want to mention is that between so-called *trivial* and *substantial* attitudes. A person's attitude is called trivial if it isn't significant and substantial if it is significant to her life or self-conception (cf. Cassam 2014; Schwitzgebel 2012). My belief whether it is raining, or my liking of strawberries or raspberries are clearly trivial attitudes, whereas my desire to have another child or the importance I attach to my family are obviously substantial attitudes.

These distinctions provide an overview of what kinds of attitudes should be envisaged in speaking of the scope of Moran's or an agential transparency account. This is not to say that such an account should necessarily apply to each of them. We should thus ask, on the one hand, what the scope of a transparency account of self-knowledge should be and whether this scope is met by a particular account. On the other hand, one may also stick to the scope delineated by the account itself and inquire whether the account actually meets its delineated scope.

In the case of Moran's account, it is explicitly specified that his account should apply, first, to intentional mental attitudes of the doxastic, conative and affective variety. Secondly, his account is supposed to hold for conscious attitudes, in the Freudian sense, of the occurrent, standing, explicit, and tacit type. It isn't supposed to hold for either implicit or unconscious, in the Freudian sense, attitudes. These latter attitudes fall in the category of alienated attitudes, because the person having such an attitude isn't in touch with the view purported by them. In Moran's view, mental attitudes that cannot be known transparently are attitudes from which the person is alienated. One of the basic claims pertaining to his account is that it doesn't apply to alienated attitudes. It remains to be seen whether Moran's account actually applies to the scope he himself envisages, and also, whether his depiction of the scope is the scope that transparency accounts should have.

[21] Cf. Moran (2001, 85).

5. Agency

Agential accounts of self-knowledge invoke a kind of agency in their explanation of what self-knowledge is and how it is achieved. It remains to be seen, however, what kind of agency that is supposed to be. In this section I contrast deliberative agency with agency manifested in holding a specific attitude. This section provides the background against which different appeals to agency that will be discussed in this dissertation are to be understood. What kind of agency is involved in transparent self-knowledge?

Moran invokes the deliberative stance to explain the involvement of agency in achieving self-knowledge. Recall that the deliberative stance is a stance from which the person recognizes her active involvement in her mental life and the connection between self-attributing a mental attitude and expressing one's commitments. Explicating the deliberative stance in this way, however, does not yet explain what that active involvement consists of and what it is for an agent to be committed. In other words, it doesn't yet explain the kind of agency that is involved. Calling it "the deliberative stance" obviously invokes connotations of deliberation, of actively seeking reasons, and making up one's mind about what to believe, want or do. Such a conception of agency would follow the now familiar convention that we are agents with respect to our mental goings-on because we can take a step back and reflectively endorse or oppose them. As Christine Korsgaard famously said in the book of her Tanner Lectures *The Sources of Normativity*:

> …our capacity to turn our attention on to our own mental activities is also a capacity to distance ourselves from them, and to call them into question. I perceive, and I find myself with a powerful impulse to believe. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse doesn't dominate me and now I have a problem. Shall I believe? Is this perception really a *reason* to believe? (Korsgaard 1996, 93)

Korsgaard draws attention to the fact that the capacity for reflection presents us with a problem, for it is no longer possible to just go with our impulses. 'Given that the person *can* either try to resist or not,' as Thomas Nagel (1996, 200; cf. Moran 2001, 142-3) comments on Korsgaard, '…anything he does will imply endorsement, permission, or disapproval from the reflective standpoint.' This means that the possibility of reflection brings with it the responsibility to take a stance: to determine whether *to* believe based on what I perceive. Beliefs (and other mental

attitudes as well as actions) of reflective creatures aren't mere results from the strongest impulses or sense impressions. By contrast, the reflective agent needs to determine for herself whether she has reason to believe something. And the general idea is that such determination matters to what will be believed. That is, by reflectively endorsing or opposing one's beliefs, a person normally forms and withdraws them. The idea is that this is what it means to make up your mind. As Moran writes,

> It is the normal expectation of the person, as well as a rational demand made upon him, that the question of what he actually does [believe] should be dependent in this way on his assessment of the [belief] and the grounds he has for it. (Moran 2001, 115)

This is thus a form of agency where a person believes certain things for reasons she recognizes as such, manifested in reflective endorsement of the belief.

However, there are several problems with this picture of mental agency. The first problem for the current context is that this view of agency presents us with an implausible view of self-knowledge. It seems overly intellectualistic that each self-attribution requires us to deliberate, reflect on our reasons, and to make up our mind. If you ask me whether I believe that Paris is the capital of France, I know the answer immediately and it would be unnecessarily laborious to actually reflect on the question whether Paris is the capital of France, to go over my reasons for thinking it is, and make up my mind about it.

Secondly, the picture doesn't actually explain the kind of agency that is exercised vis-à-vis one's mental attitudes. As Matthew Boyle (2011a, 3-4) analyzes, our normal vocabulary of decision, choice or voluntary action seems to be inept to capture what the right sort of agency is here: a person doesn't "step back" and survey a set of options of beliefs and then choose one. Rather, the agent reflectively endorses a belief by recognizing the cogency of some reason in favor of it. But if we ask what it is to recognize the cogency of reasons, we seem to run into the same question about agency. It is not to step back and survey a set of reasons and pick one. It is doing something else. Explicating it as recognizing the cogency of one set of reasons and denying that of others is not yet *explaining* but merely *specifying* the form of agency we seek to understand.

Moreover, such a form of agency would remain entirely *external* to the mental goings-on, as if a person who makes up her mind changes her beliefs in the same way as she would change the furniture in her room. As if a person reflectively

endorses a belief 'in the hopes of inducing' in herself the belief.[22] As Pamela Hieronymi (2009, 157) puts it: 'Exercising reflective control over one's own mind is not like surveying and tinkering under one's own hood.' One might engage in such manipulation or management of one's own mind, but it would be no different from manipulating the mind of someone else. And so, in order to do justice to the special relation of a person to her own mental attitudes and maintain that this relation is essentially different from a third-person relation, we need something else for agency to mean.[23]

It is useful here to come back to the distinction between intentional mental attitudes and sensations. Recall that Moran's account of self-knowledge is only supposed to apply to the former. One way to explicate the difference that I haven't mentioned yet is one that sheds light on the kind of mental agency that we are after. Moran often uses the expression that my intentional mental attitudes are "my own" or "up to me" in a way that my sensations aren't. In one sense of the term, if I suffer from a headache, that headache is *mine*. But in another sense, I remain passive with respect to it and it is something that just happens to me. My beliefs and other attitudes do not just happen to me in the way a headache does. Rather, they reflect my stance on how things stand in the world at large. This is at the core of my active relation to them, not that I can manipulate them (as I could do with my headache), but that these attitudes are expressive of my view of the world at large – of my various and evolving relations to my environment.[24]

Importantly, I may either respect or fail to respect this relation between my attitudes and my "evolving relations to the environment" by taking responsibility for my mental attitudes: not in the sense that they are under my voluntary control, but by committing myself to the views purported by the attitudes. I may take responsibility for them in the same way as I may take responsibility for the conclusion of my reasoning, or for the love I feel for someone, not because I could reason in whatever way I wish or love whomever I favor, but precisely because the

---

[22] Cf. Boyle (2009a; 2011a); Moran (2001, 118-19). See also fn. 6 (Moran 2001, p. 119) where Moran cites Dennett (in disagreement): 'Acting on a second-order desire, doing something to bring it about that one acquires a first-order desire, is acting upon oneself just as one would act upon another person: one schools oneself, one offers oneself persuasions, arguments, threats, bribes, in the hopes of inducing in oneself the first-order desire. One's stance toward oneself *and access to oneself* in these cases is essentially the same as one's stand outward and access to another' (Dennett 1978, 284-85).

[23] And perhaps, we also need another name for the "deliberative stance," since it isn't deliberative agency that is invoked in that stance.

[24] This is also why Moran rather speaks of *authority* than *control* (2001, 139). Moran sides here with the philosophical tradition of "self-consciousness as reflection" – a tradition most closely related to Kant, but also to Sartre who is strongly represented in Moran's views. This tradition is also of main interest to, for instance, Christine Korsgaard (1996; 2009) and Matthew Boyle (2005; 2009a; 2011a; 2019).

reasoning and love are expressive of *my own stance* (cf. Moran 2008). Understanding what it is to be active with respect to a mental attitude of mine thus means understanding what it is to be committed to the grasp of the world purported by the attitude in question. And this means that mental agency cannot reside only in forming or changing one's mental attitudes but is also manifested in having a mental attitude.[25] In the end, this is the kind of agency that should be incorporated in an account of self-knowledge. It remains to be seen how this is possible.

6. The underlying philosophical approach

The final theme that I would like to introduce is what is now known as *analytic Aristotelianism* (Thompson 2008). This philosophical approach isn't adopted throughout this dissertation, but in the course of my research, I have come to see it as crucial to understanding the nature of reasoning and the nature of transparent self-knowledge.[26] It is therefore useful to give a short description of the method here.

The best-known exemplar of this tradition is Anscombe's (1957) monograph *Intention*. One of Anscombe's key claims is that 'the term "intentional" has reference to a *form* of description of events' (1957, §47). In Anscombe's view, the form of intentional action is that it is done for a reason. Importantly, Anscombe explicitly denies that we should understand this in terms of a specific feature or property, or of stating necessary and sufficient conditions.[27] After all, *done for a reason* is not giving us any more information about what an intentional action is. For, if we want to understand what *kind* of reason we mean, and distinguish it from a mere causal reason, then we need to presuppose the same distinction that we are trying to understand. Hence, 'we should be going round in circles,' as Anscombe (1957, §5) writes. This means that Anscombe's reference to "form" actually discloses a completely different approach to intentional action: not describing a property of intentional action but its *form*.

Admittedly, this still sounds rather puzzling. So, let me try, in very brief terms, to give more methodological context to Anscombe's *Intention*. Anscombe's approach is now known as analytic Aristotelianism. The core commitment of analytic Aristotelianism is that some concepts require philosophical analysis, not in terms of

[25] cf. Moran (2001, 77); Boyle (2015, 341). The mental agency that Moran seeks to address is further developed in Boyle (2011a); Hieronymi (2009); and Moran (2008; 2012).

[26] It has also helped me to see the parallels between discussions on transparent self-knowledge, intentional action and reasoning.

[27] For recent illuminating papers on Anscombe's method, see Ford (2015); Frey (2013); Hlobil & Nieswandt (2016); Vogler (2001).

smaller parts, essential features or necessary and sufficient conditions, but rather in terms of their *logical form* – in terms of distinct categories or modes of being (cf. Boyle 2005; Hlobil & Nieswandt 2016; Thompson 2008). The logical form of a concept consists of the form of thought or the form of judgment that underlies the concept and it refers to what can be predicated of the thing in question. The logical form (or *structure*) of, say, X is revealed by analyzing the things that can be said or asked about X, and thus by analyzing our practices and abilities regarding X.[28]

A key motivation for using analytic Aristotelianism is that it helps to understand one of Moran's core claims about self-knowledge. Moran maintains that transparent or first-person self-knowledge is *categorically* different from non-transparent self-knowledge or knowledge of another person's mental attitudes. For instance, he writes that

> …for a range of central cases, whatever knowledge of *oneself* may be, it is a very different thing from the knowledge of others, categorically different in kind and manner, different in consequences, and with its own distinguishing and constraining possibilities for success and failure. (Moran 2001, xxxi)

Understanding what Moran signifies in saying that the first-personal self-knowledge that he is after is a distinct *category* ultimately requires understanding the underlying philosophical approach. A difficulty in understanding Moran's philosophical approach is that he doesn't explicate which philosophical approach he adopts, nor what kind of approach it is supposed to be. My proposal is that Moran's account and the *kind* of agency, the *kind* of rational human capacities and the *kind* of knowledge at issue make sense in a broadly Anscombian philosophical program. In other words, my suggestion is that the concept of transparent self-knowledge as advanced in this dissertation, and Moran's account in particular, make most sense when one adopts the analytic Aristotelian approach.

7. Dissertation overview

Chapter 1 begins by addressing the question what Moran's transparency claim regarding belief precisely consists of.[29] In his depictions of transparency, Moran

[28] The approach will be further addressed in Chapter 3 and the Concluding Reflections.

[29] In Moran's account, the paradigm case for transparency is belief. To arrive at a charitable evaluation of Moran, the case of belief is the place to start. This chapter is co-written with René van Woudenberg.

stays close to Evans' characterization of transparency. As Moran (2012, 212) writes, the idea of transparency is that 'a person can answer a question about her own belief by addressing herself to the corresponding question about the topic of that very belief.' But how should this latter question be answered? And does the claim apply to all kinds of belief? As to the first question, it seems that Moran's writings support three different kind of claims: that there aren't any conditions on how to answer that question, that one should refer to reasons in favor of the proposition believed, and that one should refer to reasons justifying the proposition believed. We evaluated these different requirements by testing whether they apply to numerous examples of beliefs, ranging from recalcitrant beliefs to beliefs based on no evidence. Based on the expositions of the examples, it is argued that Transparency is most plausible, i.e., has the widest scope, if it demands the least on how the question about what one believes is answered. But there remains a reservation about this most plausible account of transparency: it seems to be disconnected from the deliberative stance and thus in tension with the motivation behind Moran's transparency account.

Chapter 2 concerns the Two Topics Problem (TTP): the problem that the truth of *p* doesn't seem to provide an epistemic basis for the truth of *I believe that p*. A careful glance at the state of the debate on transparent self-knowledge shows that there is no consensus of what the relation between *p* and *I believe that p* might be, nor what kind of solution respects the commitments of transparency views. The main aim of this paper is to make TTP "transparent": to provide a grasp of the nature of the different responses to TTP. The responses that I will discuss are: 1) the view that TTP is only apparent; 2) inferential views; 3) judgment views; and 4) metaphysical views. In very general terms, the proceeding arguments are as follows. First, I argue that TTP has to be accepted as a genuine problem insofar as one accepts the (transparency)[30] intuitions that in self-ascribing a belief that *p* a person both makes an empirical claim that she is in a certain state of mind and endorses *p*. Secondly, taking Byrne's account as exemplary for the inferentialist response, I contend that a crucial assumption in his account, namely that inference from a premise entails belief in that premise, is unwarranted (a claim that is corroborated in Chapter 3). Thirdly, I will argue that, albeit for different reasons, both the judgment views and the metaphysical views need to presuppose a form of attitudinal awareness, i.e., an awareness of one's judgment or belief regarding *p*. It might be that a transparency account of self-knowledge should comprise an attitudinal form of awareness, but

[30] "Transparency" is in brackets, because the intuition is shared, not only by proponents of transparency accounts of self-knowledge, but also by some of those proposing different accounts of self-knowledge. Cf. Finkelstein (2003); C. Peacocke (1998).

then we need an explanation of why this would still be a transparency account of self-knowledge.

Chapter 3 develops an argument against the claim that *all* reasoning necessarily involves a change in attitudes (and thus involves attitudes regarding the premises and conclusion). Although it seems obvious that reasoning often involves such a change in attitudes, e.g. forming, revising or withdrawing a belief, that doesn't imply that a change in attitudes is necessarily involved in reasoning. For instance, we quite often reason hypothetically or merely check the validity of an argument, without having determined for ourselves whether we believe the premises. As Wright (2014, 28) has put it, we should 'distinguish inference in general from *coming to a conclusion*…; no particular attitude to [a] proposition is implicit in inference itself.' By discussing examples of reasoning without a change in view, it will become clear that a different approach to reasoning is needed: namely, one that includes instances of reasoning with and without change in attitudes. By combining insights from Anscombe and Frege, I will propose an analytic Aristotelian alternative view of reasoning, i.e., what I call the *form view*, which holds that when a person reasons she (1) makes use of conditionals, manifested in (2) a judgment of the form *p as following from q*. The corollaries of this view are that neither mental attitudes nor personal-level mental processes are necessarily involved in reasoning.

Chapter 4 takes up the question how Moran's account could be translated to mental attitudes other than belief. It assumes that Moran's transparency account works for belief and then seeks to apply it to emotion. The basic difficulty in such application is that where the relevant "outward-directed" question for belief is simply whether the proposition under consideration is true, the relevant "outward-directed" question for emotion is less easy to discern. The reason for this is that emotions do not only seem to be about the world, but also about what is important to the person having the emotion. Even if we all agree that a person has betrayed me, I need not *feel* betrayed if either the person or the betrayal itself were insignificant to me. Similarly, only if I care about a sports team, will their wins and losses spark joy or disappointment, respectively. We only feel an emotion if something matters to us (cf. Helm 2010). Chapter 4 thus argues that Moran's transparency claims cannot be applied to emotions, at least not without incorporating an account of the relation between transparency and what matters to us.

Chapter 5 takes up this latter question: does Moran's account apply to substantial mental attitudes, such as one's "cares", concerns, and values? One might think that

whereas perhaps trivial self-knowledge is a result of the special relation a person has to her own mental life as portrayed by Moran – especially, that she is in a position to avow her mental attitudes – substantial self-knowledge is not. Were a person, say Katherine, to desire another child, this should be reflected not only in her avowal on the matter but also in a wide range of actions and reactions (cf. Lawlor 2009). In this chapter I take up this challenge and argue to the contrary: even if such patterns of action and reaction form part of coming to know my substantial mental attitudes, avowing these attitudes remains essential and has a unique status in coming to know them. My arguments show that the status of avowal is unique, first, because the significance of patterns of action and reaction, and what such patterns tell about our attitudes, ultimately depends on avowal. And secondly, because substantial mental attitudes require one to have a self-conception. Acquiring self-knowledge of substantial mental attitudes is a struggle to fulfill the commitments pertaining to these attitudes. And it is a struggle that requires a person to manifest her agency – to take responsibility for who she is and putting herself at risk of being challenged and making mistakes. Or so I will argue.

CHAPTER ONE

Three Transparency Principles Examined[1]

Abstract

This paper derives, from Richard Moran's work, three different accounts of doxastic Transparency – roughly, the view that when a rational person wants to know whether she believes that *p*, she directs her attention to the truth-value of *p*, not to the mental attitude she has vis-à-vis *p*. We investigate which of these is the most plausible of the three by discussing several (classes of) examples. We conclude that the most plausible account of Transparency is in tension with the motivation behind Transparency accounts: it is disconnected from the deliberative stance.

1. Introduction

A widely discussed phenomenon in the philosophy of self-knowledge is Transparency. The most central case is *doxastic* Transparency: when I answer a question about whether I believe that *p*, I attend to the same things I would attend to if I were answering the question whether *p* is true (see Edgley 1969, 90; Evans 1982, 225).[2] A prominent recent proponent of such an account is Richard Moran

[1] This chapter is co-authored with René van Woudenberg and will be published as: Woudenberg, R. van and N. Kloosterboer. 2019 (forthcoming). "Three Transparency Principles Examined." *Journal of Philosophical Research* 44.

[2] Mental attitudes other than belief, for example desires, have been argued to be transparent too (see, for example, Fernandez 2013, ch. 3). We focus on doxastic Transparency.

(2001). His views have proved hard to pin down, however: what exactly does he take Transparency to be[3], and what is its scope?[4]

In this paper, we take up these questions by taking a fresh and analytical look at Moran's work on Transparency. Our aim, however, is not to determine Moran's views, but rather to explore a number of suggestions that can be found in his work as to the nature of Transparency. Our second aim is to investigate the scope of these accounts: are there (classes of) beliefs that don't conform to Transparency? Hence, this is not a study in Moran-exegesis (although, of course, we engage in some exegesis) but a study of what will turn out to be three accounts of Transparency that are suggested in his work. This, then, is the task for this paper – a task that, to the best of our knowledge, no one else has undertaken.

The paper is organized as follows. In section 2, we sketch the motivations behind Transparency accounts of self-knowledge. Section 3 presents a list of Moran-quotations from which we derive three different Transparency Requirements, as we shall call them. Section 4 examines whether these Requirements are true by discussing (classes of) cases. Section 5 concludes that the most plausible of the three Transparency Requirements is in tension with the motivation behind Transparency.

2. The motivation behind transparency accounts of self-knowledge

An important motivation behind Transparency accounts of self-knowledge is that there seem to be a number of asymmetries, epistemic as well as non-epistemic[5], between how we know our own minds (for instance how we know what we believe, hope, etc.) and how we know the minds of others. Knowing one's own mind seems to have an authority, directness and non-evidential basis that one's knowledge of other minds lacks. Moran accounts for these asymmetries by distinguishing two

[3] As we shall see in section 3, different Transparency accounts can be derived from Moran (2001). Moreover, the literature on Transparency provides descriptions of Transparency that differ from the ones that we distil from Moran (2001), for instance Byrne (2005), Gertler (2012), Finkelstein (2012), Silins (2012), Fernandez (2013), Cassam (2014), Barnett (2016).

[4] One point of criticism raised against Moran is that, on his account(s), Transparency's scope is restricted to *rational* mental states and hence in the case of beliefs to *rational* beliefs only. According to the critics, his account(s) do(es) not explain how we know, as we sometimes do, that we have irrational beliefs, which in turn means that there are more ways to obtain self-knowledge than through Transparency. Criticisms along these lines are offered by Lawlor (2009), Finkelstein (2012), Paul (2012), and Cassam (2014). As we will explain further on, Moran holds that Transparency doesn't apply to irrational beliefs, but only if they are irrational because they are *alienated*.

[5] The more traditional accounts of self-knowledge, especially the so-called perceptual view, focused almost exclusively on the epistemic asymmetries between knowing one's own and knowing someone else's mind.

"stances" that one may have vis-à-vis one's own mind, the "first-person" or "deliberative" stance, and the "third-person" or "theoretical" stance. These stances correspond with two kinds of questions one may pose about one's own mental life. A theoretical question about one's mental life is 'one that is answered by discovery of the fact of which one was ignorant,' whereas a deliberative question is one that 'is answered by a decision or commitment of some sort, and it is not a response to ignorance of some antecedent fact about oneself' (2001, 58).

When one answers the question "Do I believe that *p*?" from a theoretical stance, one attends to the sort of evidence that could, in principle, also be consulted when *someone else* inquires whether I believe that *p* – evidence consisting in one's behavior and one's utterances. But when one answers the question from the deliberative stance, one doesn't look for evidence about one's own behavior and one's own utterances, instead one reflects on whether *to* believe that *p*. Successful deliberative reflection terminates in *forming* an attitude, so in believing that *p*.

Moran construes a close relation between Transparency and the deliberative stance:

> the vehicle of transparency … lies in the requirement that I address myself to the question of my state of mind in a *deliberative* spirit, deciding and declaring myself on the matter, and not confront the question as a purely psychological one about the beliefs of someone who happens to be me. (2001, 63)

Transparency requires that one answers "Do I believe that *p*?" from the deliberative stance. It requires that one concentrates on *p* (and answers "Is *p* true?") rather than on the psychological attitude one has vis-à-vis *p* (one should *not* set oneself to answer the question "Do I have the belief-attitude towards *p*?"). Transparency requires that ascribing to oneself the belief that *p* is connected to taking *p* to be true. This is related to the connection that Moran draws between Transparency and Moore's paradox, according to which it is paradoxical to assert "*p*, but I don't believe it" or "I believe that *p*, but not-*p*" (2001, 68-73, 83-84). From the theoretical stance, such assertions seem perfectly sound, because the self-attribution of the belief that *p* is unconnected to one's endorsement of *p*'s truth. From the deliberative stance, by contrast, uttering (or thinking) Moore-type sentences is paradoxical if not plain irrational. Given that Moore-sentences are indeed paradoxical and irrational, the fact that Moore-type sentences are sound from the theoretical stance but paradoxical and irrational from the deliberative stance speaks in favor of the deliberative stance. In fact, it is one of the main reasons for Moran to say that

rationality requires a person to answer questions about her state of mind from the deliberative stance (and thus through Transparency).

Transparency doesn't apply to alienated beliefs, according to Moran (2001, 85-93). A belief is alienated when: one lacks reasons in support of its content; one cannot make sense of the fact that one has it; and it fails to link up with other beliefs one has. Examples of alienated beliefs include repressed beliefs, unconscious beliefs, and delusional beliefs. We propose to put this as follows:

> *Alienated*: One's belief B is alienated=$_{df.}$ B persists independently of one's reflections and criticisms of B and one cannot make sense of one's having B.

Moran subscribes, we think, to the following biconditional:

> Biconditional: (a) If one's belief is transparent, it isn't alienated; and (b) if one's belief isn't transparent, it is alienated. Which is to say: one's belief is transparent iff it isn't alienated.

Connecting this Biconditional to Moran's statements about rationality yields the following: Rationality requires that one answers "Do I believe that *p*?" in a transparent way (Moran 2001, 84, 93, 107-13), which, per Biconditional, cannot be done if one's belief is alienated. If an alienated belief prevents one from satisfying a rationality requirement, then such a belief is irrational.[6] Moran, we therefore say, subscribes to:

> Conditional: If one's belief is alienated, then it is irrational.

Summing up, the motivation behind Transparency is to capture essential aspects of the first-person stance vis-à-vis one's own beliefs. According to Moran, these essential aspects are manifest in the deliberative stance. But *when* is belief transparent: which conditions must be satisfied for one's belief to be transparent? That is the topic of the next section.

[6] We assume here that if a *person* is irrational due to not forming the belief that *p* in a Transparent way, *the belief that p itself* is thereby rendered irrational.

3. Three formulations of transparency

Consider the following quotations from Moran:

> Quotation 1: Ordinarily, if a person asks himself the question "Do I believe that P?," he will treat this much as he would a corresponding question that does not refer to him at all, namely, the question "Is P true?" And this is not how he will normally relate himself to the question of what someone else believes. (2001, 60)

> Quotation 2: [F]rom within the first-person perspective, I treat the question of my belief about P as equivalent to the question of the truth of P. (2001, 62-3)

> Quotation 3: [A] person can answer a question about her own belief by addressing herself to the corresponding question about the topic of that very belief. (2012, 212)

> Quotation 4: When asked 'Do I believe P?', I can answer this question by considering the reasons in favour of P itself. (2003, 405)

> Quotation 5: [T]he claim… is that a first-person present-tense question about one's belief is answered by reference to (or consideration of) the same reasons that would justify an answer to the corresponding question about the world. (2001, 62)

These quotations state claims about what subjects do, or can do, when answering the question "Do I believe that *p*?" When paired with what Moran says about the relationship between Transparency and rationality, these claims have normative implications as well.[7]

Transparency can thus be thought of as a claim about how a *rational* person S, from her first-person perspective, answers the question "Do I believe that *p*?" (Let us call this last question "QBelief", or QB for short.) Alternatively, it can also be thought of as a claim about how one rationally *ought* to answer QB.

Now *how* ought we go about answering QB? Quotation 1 says that a rational person "will treat" QB as she would treat the very different question "Is *p* true?" (Let

[7] See also Moran 2001, 62-63, 84.

us call this latter question "QProposition", or QP for short).[8] Alternatively, it is, as quotation 2 has it, that a rational person "treats" QB as "equivalent to" QP. And quotation 3 says that a person answers QB by answering QP. These quotations, we take it, say essentially the same thing. Quotation 4 says something different, viz. that when we answer QB, we answer it by "considering the reasons for P itself". Quotation 5 says something different yet, viz. that the answer to QB must be given "by reference to the same reasons that would justify" an answer to QP. So, the quotations suggest three different requirements for conforming to Transparency when answering a QB question (we call these requirements for Transparency TR_i):

> TR_1- one answers QB by answering QP;
> TR_2- one answers QB by considering the reasons in favor of *p* itself (=by considering the reasons relevant to answering QP);
> TR_3- one answers QB by reference to the same reasons that would justify an answer to QP.

These three ways to adhere to Transparency are distinct as they put increasingly stronger requirements on how to answer QB questions. When filled out the first way (TR_1), no reference to reasons is required, while the other ways do require such reference. The difference between way TR_2 and TR_3 is that whereas the former requires "reasons in favor of P", the latter requires that those reasons "justify" an answer to QP. The difference is that while reason R may be a reason in favor of P, R may be insufficient to "justify" an affirmative answer to QP. (That the last game went well, may be a reason for thinking that the next game will go well too, but it may be insufficient for an affirmative answer to the question "Will the next game go well?") Moran nowhere explicitly states what justification requires. In what follows we will work with the notion that proposition P is a reason, R, in favor of proposition P*, iff P raises the probability of P* (i.e., $Pr(P^*|P) > Pr(P^*)$). And we shall work with the notion that a reason R justifies proposition P iff (i) R raises the probability of P enough to make it rationally permissible to believe that P, and (ii) R's probability is such that it is rationally permissible to believe that R (i.e. R is credible).

[8] QB and QP are very different indeed: QB is about what someone believes, whereas QP is about the truth of a proposition; QB makes an essential reference to a first-person perspective (see Edgley 1969, 90), QP does not; QB is "inward-looking", whereas QP is "outward-directed" (to use Evans' phrase; see Evans 1982, 225). Moreover, the truth conditions for the answers to QB and QP are different.

4. Transparency requirements examined

We now consider a number of beliefs, each of which is paradigmatic for a large class of beliefs and discuss whether they satisfy Transparency in one of its three formulations.

A. Obsessive beliefs

Some people have obsessive beliefs. Consider Jane who believes that she will fail the exam, even though she is aware of many reasons for thinking that she can pass it, and no reason whatsoever for thinking she will fail. She knows she has prepared diligently, and also that when she prepared for examinations in this way in the past, she passed with flying colors. Yet she believes, and knows she believes, that she will fail. The only reasons she is aware of are in support of the proposition that she will pass the exam, but those reasons are disconnected from her belief. Still, if she reflects on the question whether she will fail the exam, she cannot but think that she will. Is Jane's belief Transparent?

Prior to addressing this, we note that Jane's belief that she will fail the exam conforms to *Alienated*, as it persists independently of the reasons in support of the proposition that she will pass the exam and is a belief that she cannot make sense of. It sits in her mind like a stone. (The only way she can make sense of it is by adopting a theoretical stance towards it and ascribe it, say, to an excessive form of anxiety.) Her belief, therefore, is alienated. But then, by Biconditional and Conditional, her belief isn't transparent, and hence is irrational.

Do the Transparency Requirements give this verdict? Let us first consider TR_3. Does Jane answer "Do I believe I will fail the exam?" by reference to the same reasons that would justify an answer to "Will I fail the exam?" She doesn't do that, for the reasons that are available to her support the proposition that she will pass the exam, yet she believes she will fail. So, Jane's belief does not conform to TR_3. Hence it isn't Transparent if that requires what TR_3 formulates. This is as it should be, since her belief is not transparent – because her reasons are opaque to her, it satisfies *Alienated*.

Jane's case isn't a counterexample to TR_2 either. For as we have described the case, Jane *doesn't* answer "Do I believe I will fail the exam?" by considering reasons relevant to the proposition that she will fail the exam. Here, too, things are as they should be.

With respect to TR_1, however, things are different. For Jane can and, let us assume, *does*, answer "Do I believe I will fail the exam?" by affirmatively answering "Will I fail the exam?" She cannot but view the matter this way, despite her reasons

to the contrary. Hence her belief conforms to TR_1. But since her belief *is* alienated, this is not what we would expect. Meeting one of the three Transparency Requirements (each of which is suggested by the Moran quotations to state both necessary and sufficient conditions for a belief to qualify as Transparent) should, as follows from Biconditional, stand and fall together with a belief not being alienated. Hence, that Jane's belief conforms to TR_1 actually counts against TR_1.

We can put the same point also as follows.[9] If TR_1 is correct (i.e., if that is how we must understand what Transparency requires), then one's obsessive beliefs will most often be Transparent. Therefore, and here is the important point, TR_1's formulation of Transparency doesn't explain why obsessive beliefs will often be alienated and irrational. So, if Transparency is understood as formulated in TR_1, we have a counterexample to Moran's remarks that suggest an intimate connection between transparency and alienation/irrationality (as captured by the Biconditional and the Conditional) and a lack of transparency. And this is a drawback for TR_1.

Since Jane's obsessive belief is alienated, we conclude that TR_3 and TR_2 give the correct verdict that Jane's belief is not Transparent, whereas TR_1 wrongly entails that Jane's belief is Transparent. This means that in this case, TR_3 and TR_2 are more adequate formulations of Transparency than TR_1.

B. Beliefs based on non-justifying reasons

Consider next a paradigm case of a belief that is based on non-justifying reasons. Several months and sometimes even several years after World War II had come to an end, some persons who had been imprisoned in concentration camps or who had been hiding in far-out places and who were widely considered as missing, would repatriate, and show up in their home towns. Suppose Elisabeth's husband had been deported during the war, and in 1949 still has not come home. Suppose further that Elisabeth nonetheless keeps on believing that her husband will return home.[10] When asked why she believes this, the reason she gives is that there have been other men who have returned home long after the war was over. She is aware of countervailing reasons such as (i) that the war is over for four years now and that it is unlikely that Nazis still keep prisoners, (ii) that many prisoners who died in the camps remained unidentified, (iii) that if her husband was alive and free, he would have come home by now. Still, her belief that her husband will return persists, as

[9] This was suggested to us by an anonymous reviewer for *Journal of Philosophical Research*.

[10] This example was developed independently from, but bears similarities to, Finkelstein's example of Lana (2003, 166).

she thinks these countervailing reasons are trumped by her positive reason.[11] Is Elisabeth's belief Transparent?

Before addressing this matter, we note two things. First, Elisabeth's belief is unjustified. Her reason itself is barely credible, and the evidence (the proposition that some men have returned home even years after the war, in conjunction with (i), (ii), and (iii)) is radically insufficient to render her belief rationally permissible. Her belief is irrational. Second, Elisabeth's belief, unlike Jane's belief in the previous example, doesn't seem to be alienated. After all, her belief is *not* disconnected from her reflections and criticisms of her belief, and she *can* make sense of her belief. Her reason to believe that her husband will return is that there have been other men who made a very belated return home after the war. Her reason for believing as she does is *not* opaque to her – she knows and cites her reason to anybody asking. So here, we note a difference between Elisabeth and Jane. In the case of Elisabeth's belief, reasons *do* play a role (albeit in a wrong way), whereas in the case of Jane's belief they don't. This suggests that while the absence of (weighing) reasons (as is the case with Jane) makes a belief alienated, wrongly weighing the reasons (as seems the case with Elisabeth) does not. So, Elisabeth's belief isn't alienated. Is it Transparent?

Well, does Elizabeth answer "Do I believe my husband will return home?" by reference to the same reasons that would justify an answer to "Is it true that my husband will return home?" As we have indicated, the reason that Elisabeth has is insufficient to justify her belief. Hence, if Transparency requires what TR_3 specifies, Elizabeth's belief is not Transparent. However, we indicated that Elizabeth's belief is *not* alienated. This state of affairs goes against the Biconditional, according to which Transparency and non-alienation stand and fall together. Hence, assuming we want to maintain the Biconditional, TR_3 is not an adequate formulation of Transparency, as it doesn't have the right entailment (it entails that Elizabeth's belief is alienated, which, we have suggested, it is *not*).

Now consider TR_2. Elisabeth *does* answer "Do I believe that my husband will return home?" by considering reasons for the proposition that her husband will return home. For she has a (weak, to be sure) positive reason for her belief and she thinks this positive reason, somehow, trumps the negative reasons (i), (ii), and (iii). So, if Transparency requires what TR_2 formulates, then Elizabeth's belief *is*

[11] Others have raised objections against Transparency. See, e.g., Barnett (2016, sec. 4.2), Silins (2012, 304-5). However, their objections target Transparency theses that are different from our TRs; Barnett's target is the thesis that "*P* is a good reason for you to believe that you believe that *p*", and Silins' target is the thesis that "If you judge that *p*, then your judgment that *p* gives you immediate fallible justification to believe that you believe that *p*."

Transparent. Given Biconditional, her belief is not alienated – which we agreed it was not. Hence, Elizabeth's belief is not a counterexample to TR_2.

It is not a counterexample to Transparency as formulated in TR_1 either. For Elisabeth can and likely does answer QB "Do I believe my husband will return home?" by answering the corresponding QP "Will my husband come home?" So, if Transparency requires the satisfaction of TR_1, Elizabeth's belief is Transparent. Given Biconditional, her belief is not alienated – which we agreed it isn't.

We conclude that Elisabeth's belief is a counterexample to Transparency, if Transparency requires the satisfaction of TR_3, but that it is not a counterexample to Transparency if it requires the satisfaction of TR_2 and TR_1.

However, TR_2 and TR_1 (in conjunction with the Conditional) do not give us the verdict that Elisabeth's belief about her husband's return is irrational. What we see here is that when a belief is irrational because it is based on non-justifying reasons and not because it is alienated, TR_2 and TR_1 don't speak to that. On the one hand, this seems to be a good thing, for we quite often have irrational beliefs and also *know* that we have them. Moreover, respecting Transparency is a necessary, not a sufficient, condition for being rational. On the other hand, however, the motivation underlying Transparency seems to harbor the hope that Elisabeth would do more than just recite her reasons or recite her commitment to the truth of *p*. The reason for this is that QB is, from the deliberative stance, seen as the question whether *to* believe that p. Therefore, the hope of the deliberative stance seems to be that Elisabeth would think twice whether she *should* be committed to the truth of *p*. This isn't meant as a flat-out objection to either TR_2 or TR_1, but as revealing a tension between Transparency and the deliberative stance – i.e., with the central motivation behind Transparency. For TR_2 and TR_1 should give us the verdict that Elizabeth's belief is irrational.

C. Long-standing beliefs

As Shah and Velleman have argued, the QB question "Do I believe that P?", can be posed in two quite different ways.[12] We may ask that question while the answer is entirely open to us. Suppose the question is "Do I believe that cricket originated in Burma?", and this is a fresh topic for us. Then, if we answer this question rationally, we will look for reliable information, weigh the evidence, and make up our minds. If the evidence that cricket originated in Burma is compelling, we will believe this. If we proceed this way, we satisfy the requirements of Transparency in each of its

[12] Shah and Velleman (2005, 506) distinguish between the deliberative question whether *to* believe that *p* and the question *whether* one believes that *p*.

three formulations. For we answer this QB by answering the corresponding QP ("Did cricket originate in Burma?"), and so satisfy requirement TR_1. But we also satisfy requirement TR_2, for we answer QB by considering the reasons in favor of *p* itself. We even satisfy requirement TR_3 for we answer QB by reference to the same reasons that would justify an answer to QP. If the question "Do I believe cricket originated in Burma?" (and this is a fresh topic for us, one about which we have no opinions yet) is paradigmatic for the large class of beliefs obtained by deliberation, we cannot expect to find counterexamples to the requirements of Transparency as formulated in TR_1, TR_2 or TR_3.

However, QB can be posed in quite another way as well. We may ask "Do I believe the greenhouse effect is real?" after having given the matter a great deal of thought and after having made up our minds about it such that it has become a long-standing belief of ours that the greenhouse effect is real. How do we go about answering a QB type question about a long-standing belief of ours?

Normally we don't go over the possibly complicated and manifold reasons we have regarding *p*. It would seem time-consuming, unpractical and even irrational if we were to do that. We know what our long-standing beliefs are; we don't have to figure that out by rehearsing and reviewing the reasons and evidence that we have collected over the years. Of course, we can try to take a fresh look at our long-standing beliefs and review our reasons and evidence that led us to form the belief in the first place. But doing this is certainly not needed if we want to find the answer to the question "Do you believe that *p*?" when we have believed *p* for a long time. We can answer such a question very quickly, without going through any laborious process.

We note that long-standing beliefs are not, as such, alienated. Long-standing beliefs, like your belief in the greenhouse effect, need not persist independently of your reflections and criticisms, and you may be perfectly capable of making sense of it. Your belief is not disconnected from other beliefs of yours, your reasons are not opaque to you – the point is only that you don't have to consult them in order for you to be able to answer QB. Of course, it is possible that one of your long-standing beliefs is alienated. But if it is, it isn't alienated *because* it is long-standing.

Are long-standing beliefs Transparent? From what we have said above about long-standing beliefs, it seems to follow that they conform to neither TR_2 nor TR_3. For at each particular point in time we almost invariably know that we have the long-standing belief that *p without* considering, at that point in time, the reasons in favor of *p* itself, and a fortiori *without* referring to or consulting reasons that would justify an answer to "Is *p* true?". This means that *at that point in time*, our belief that

p is not Transparent.[13] But if a belief is not Transparent, then by Biconditional and Conditional it is alienated and irrational. We have suggested, however, that long-standing beliefs are not, as such, alienated or irrational. From this it follows that long-standing beliefs are counterexamples to Transparency, if that requires the satisfaction of TR_2 and TR_3.

Long-standing beliefs may still be Transparent, if it requires what TR_1 formulates. If your belief in the greenhouse effect is long-standing, do you then answer "Do I believe in the greenhouse effect?" by answering the question "Is the greenhouse effect real?"? You very well may. Note that proceeding this way isn't laborious; it doesn't require you to have reasons, let alone to produce justifying reasons for the reality of the greenhouse effect. It merely requires that you take the greenhouse effect to be real.

We conclude that long-standing beliefs only conform to Transparency if it requires what TR_1 formulates, not what TR_2 or TR_3 formulate. Since long-standing beliefs are neither alienated nor irrational, they should conform to Transparency and so TR_1 is the proper formulation of it in the case of long-standing beliefs.

D. Basic beliefs

Foundationalists have argued that some of the things we know are "basic" in the specific sense that the belief that constitutes the knowledge is "immediately justified," by which they mean, roughly, that the justification of the belief does not derive from, nor is it based on, other beliefs that the subject has in favor of the belief or that constitutes the evidence for it.[14] Foundationalists hold that some basic beliefs are "properly" basic, whereas others are not. "Properly" basic beliefs have warrant, or justification or some other positive epistemic status – something that "improper" basic beliefs lack. Paradigmatic examples of "properly" basic beliefs are belief in self-evident propositions, such as the Principle of Identity, A=A, and the proposition that 2+1=3. Many also take belief in incorrigible reports from experience such as "I seem to be seeing something blue", to be "properly" basic.

---

[13] One might think that one's long-standing beliefs do *not* violate TR_2 and TR_3 when or because the reasons one once had can be recalled (even when in fact they are not actually recalled). However, when one knows one believes that *p*, and one does not actually recall at that moment one's reasons for *p*, one knows what one believes at that moment in a way that is *not* transparent – one's knowledge just isn't based on reasons pertaining to *p*. But believing that *p* is not, on account of its non-transparency, alienated. (Thanks to an anonymous reviewer for pressing us on this matter.)

[14] Versions of foundationalism have been developed by Chisholm (1982), Audi (1998) and Plantinga (1993). For a defense of the idea that some beliefs are basic against several traditional objections, see Van Woudenberg (2005).

Let us now assume that there *are* properly basic beliefs and let us also assume that a paradigm case is Rik's belief in the Principle of Identity, A=A. Rik's belief is not alienated. He can make perfect sense of it, and when he reflects on the Principle, he feels compelled to believe it, even though he has no, and can provide no argument that has the Principle as its conclusion, and whose premises are more evident than the Principle itself. That is to say, his belief in the Principle is justified (has positive epistemic status – even quite a lot of it) even though the justification is not provided by arguments that Rik has, and that constitute the evidence for it. Moreover, his belief links up perfectly with other beliefs that he has, for instance with his belief that identity is transitive, reflexive and symmetrical, and with his belief that the Evening Star is the Morning Star. Nor does his belief float free from what he thinks is the truth about the Principle, etc. We also note that Rik's belief doesn't look irrational. Rather, it seems Rik cannot be rational unless he accepts the Principle of Identity.

Is Rik's belief Transparent? Given what we have said above, it readily follows that it conforms to neither TR_2 nor TR_3. When Rik asks "Do I believe the Principle of Identity?" he does not answer this question by reference to reasons, and a fortiori not by reference to reasons that would justify an answer to the question "Is the Principle of Identity true?" So, his belief is not Transparent, if Transparency requires what TR_2 or TR_3 formulate. But by Biconditional and Conditional this entails that Rik's belief is alienated and hence irrational. But this, as we suggested (and as foundationalists will agree), is wrong. This means that Rik's belief is a counterexample to Transparency, if Transparency requires conformity to TR_2 and TR_3.

With respect to TR_1 things are different. For Rik *can* and, in all likelihood, *does* answer "Do I believe the Principle of Identity?" by answering "Is it true?" And his answer to the latter question is easy enough, as the Principle is overwhelmingly plausible for him, as plausible as any self-evident proposition is: the moment he realizes what the Principle says, he sees that it is true. So, Rik's belief is Transparent, if that requires conformity to TR_1. Biconditional and Conditional entail that Rik's belief is neither alienated nor irrational. Which is how it should be.

We conclude that, on the assumption of foundationalism, properly basic beliefs are not Transparent, if Transparency requires what TR_2 and TR_3 formulate. We also conclude that if foundationalism as described is correct, and if properly basic beliefs are to be Transparent, TR_1 must be deemed the proper formulation of its requirement.

E. Belief in anti-skeptical propositions

Consider next the class of "anti-skeptical" propositions, such as *There is an external world* and *I am not a brain in a vat (BIV)*. A peculiarity about belief in anti-skeptical propositions that has often been noted is that there is a simple and seemingly persuasive argument for the conclusion that it can never amount to knowledge. The argument is the famous closure-based argument for radical skepticism.[15] The first premise of the argument is that if one is to have knowledge of a wide range of everyday propositions (call them OHs), then one must be able to rule out radical skeptical hypotheses (such as the BIV-hypothesis) that one knows to be incompatible with the OHs. The second premise is that we cannot rule out the skeptical hypotheses. From which it follows by modus tollens that one cannot know OHs.

For our purposes, it is relevant to see why the second premise is generally held to be true. It is held to be true because, as the history of skepticism seems to make clear, all the evidence that we have is compatible with both the skeptical propositions and their denials.[16] As Robert Nozick once put it: 'If one of these other things [i.e. skeptical scenarios] was happening, your experience would be exactly the same as it is now. So how can you know none of them is happening?' (Nozick 1981, 167) To be sure, most philosophers, and virtually all non-philosophers, believe the anti-skeptical propositions. Such beliefs are paradigmatic examples of what Thomas Reid (1764 [1997], 170) has called common sense or "instinctive" beliefs, and what Ernest Sosa (2009, 21) called "animal beliefs". We cannot help having them, and even the strongest arguments seem incapable of shaking them out of us. Still, as has been argued, no reason favors the anti-skeptical propositions over their denials.[17] If this argument is compelling, it would seem to leave belief in these propositions in a fundamental sense unjustified.

Yet, belief in anti-skeptical propositions is not alienated. It is very different from Jane's belief that she will fail her examination even though all the reasons available to her point in another direction; Jane cannot make sense of her belief. But believers in anti-skeptical propositions *can* make sense of their beliefs and they do link up with many other beliefs that they have. Such beliefs are by no means irrational. Rather, they seem entirely rational, even if their rationality is not derived

[15] See Pritchard (2005, 27-8).

[16] See Machuca and Reed (2018).

[17] As is forcefully argued by Stroud (2002, 99-121). Externalists, of course, disagree with Stroud's assessment. However, Moran is certainly no partisan to externalism, and hence for a discussion of Transparency views inspired by Moran, externalist responses to the skeptical problem are not relevant. Pritchard (2005) is an in-depth discussion of the dialectic between internalist and externalist responses to skepticism.

from reasons that constitute the evidence for them. They are rational in the sense in which, according to foundationalism, properly basic beliefs are rational.

Is belief in anti-skeptical hypotheses Transparent? For concreteness sake we focus on Geraldine, who believes she is not a BIV. Suppose she asks "Do I believe I am not a BIV?" Does she proceed by answering along the lines of TR_3? What we have said about skeptical hypotheses suggests she does not, as there are no reasons (as skeptics would allow) that would justify (as skeptics don't allow) an affirmative answer to "Am I a BIV?" This means that Geraldine's belief is not Transparent, if that requires the satisfaction of TR_3. By Biconditional and Conditional it follows that Geraldine's belief is alienated and irrational, which we have suggested it is not.

But if Transparency requires what TR_2 specifies, then Geraldine's belief may very well be Transparent. For she may answer the question "Do I believe I am not a BIV?" by considering reasons in favor of the proposition that she is not a BIV, even if those reasons are insufficient for *justifying* belief in it. What could such reasons be? They might include that here is a hand, and there another one; that it surely looks like there is an external world; that there is no reason to think that she is a BIV, etc. So, if Transparency requires what TR_2 specifies, Geraldine's belief *is* Transparent. And by Biconditional and Conditional it is neither alienated nor irrational, which is as it should be.

Now if Geraldine's belief is Transparent where that requires the satisfaction of TR_2, then a fortiori it is also Transparent if that requires the satisfaction of TR_1. For, as we indicated at the end of section 3, the three Transparency Requirements put increasingly stronger demands on how to answer the QB question.

The conclusion of this section, therefore, is that if Transparency is to hold for anti-skeptical beliefs, TR_3 is not the proper formulation of its requirement.

F. Cases of forgotten evidence and testimonial beliefs

Casimir once had an excellent reason for believing that the square root of 2 is not a fraction – the reason being that he worked through the proof himself. Later on in life, alas, he forgot how the proof went, but still remembers the theorem and continues to believe it. This is a paradigm case of a belief based on forgotten evidence.[18]

Related to this is the case of someone who, like Casimir, believes the theorem, but unlike Casimir not because he once worked through the proof himself, but because of the testimony of others. Virtually all people who have heard of Andrew Wiles' proof of Fermat's so-called Last Theorem, are in this position vis-à-vis the

[18] For discussion of similar cases, see Boyle (2019, 5-6) and Byrne (2005, 84-5).

proposition that $x^n+y^n=z^n$ has no solutions for n>2. Let us suppose that Agnes is in such a position: she believes this proposition on the basis of testimony. Agnes' belief too is paradigmatic for a very large class of beliefs, namely beliefs based on testimony. We consider both cases together.

First, let's consider whether Casimir's and Agnes' beliefs are alienated. Both can make sense of their beliefs. Hence, it would be wrong to say that their beliefs are alienated. It would also be mistaken to say that their beliefs are irrational. It has been argued that belief for which one has forgotten the evidence, can still be rational.[19] And most epistemologists agree that belief based on testimony can be fully rational.[20]

Casimir's and Agnes' beliefs thus aren't alienated nor irrational, but are their beliefs Transparent? We turn to Casimir's belief first. Two evaluations of his case suggest themselves. The first goes as follows. When Casimir asks QB: "Do I believe this theorem?" he cannot answer it by reference to the same reasons that would justify an answer to QP: "Is this theorem true?" After all, he has forgotten the proof, and so lacks reasons that can justify an answer to QP. This means that his belief is not Transparent if that requires the satisfaction of TR_3. By Biconditional and Conditional, this entails that his belief is alienated and irrational. But this, as indicated, is a claim that is difficult to maintain. Hence Casimir's belief is a counterexample to Transparency if TR_3 is the proper formulation of it. Casimir's belief is also a counterexample to Transparency if that requires fulfillment of TR_2. For Casimir doesn't, later in life, answer QB by considering reasons for the mathematical theorem, as he has forgotten the proof. Still, his later belief in the theorem is by no means irrational or alienated. Hence Casimir's belief is also a counterexample to Transparency if it requires the satisfaction of TR_2. But it is not a counterexample to TR_1. For Casimir can honestly say "yes" to the QP question "Is the square root of 2 not a rational number?", and thereby answer QB, without a proof of the theorem available. We conclude that Casimir's belief is Transparent if that requires conformity to TR_1.

However, there is a second evaluation that commends itself. Casimir has a reason for his belief that the square root of 2 is not a rational number. His reason is not the proof, which he has forgotten, but his *remembering* that he once worked out the proof for himself. When Casimir asks QB "Do I believe that the square root of 2 is not a fraction?" he may answer by reference to his remembering that he once worked out the proof himself. And by doing so, he satisfies the requirement of TR_2,

[19] For example, Foley (2001).

[20] For an overview of positions that endorse this, see Gelfert (2014).

for he answers QB by considering his reasons in favor of the theorem itself – his reason being his remembering that he once worked out the proof for himself. By doing so he also satisfies the requirements of TR_3, for his answer to QB refers to the same reason that justifies an answer to QP, "Is the square root of 2 a fraction?" To spell it out fully: Casimir answers QB "Do I believe that the square root is not a fraction?" by reference to the reason that he once worked out the proof for himself, and that same reason justifies his (affirmative) answer to QP "Is the square root of 2 not a fraction?" So, Casimir's belief is Transparent, if that requires the satisfaction of TR_3 (and hence also if that requires the satisfaction of TR_2 or TR_1).

Should one of these evaluations be preferred over the other? In order to discuss this, we introduce a distinction between "direct" reasons and "indirect" reasons. Your remembering what the proof of *p* is, is for you a *direct* reason for *p*. By contrast, your remembering that there is a proof of *p* without remembering what the proof is, is for you an *indirect* reason for *p*. The difference is that whereas a direct reason for *p* enables one to "see" for oneself, in some sense of "seeing", that *p* is true or likely to be true, an indirect reason does not enable one to "see" that. A direct reason for *p* is a reason that explains why *p* is true, or likely to be true, whereas an indirect reason does not. That is why we normally only use direct reasons for *p* if we attempt to convince someone who is skeptical about *p*, i.e. someone who says she will accept *p* only if she can "see" for herself that *p* is true or likely to be true; we don't use indirect reasons to that end.

Casimir's initial reason to believe the theorem was a direct reason – the reason being the proof that he was able to make. But after having forgotten the proof, his reason to believe the theorem was an indirect one. This is relevant to the issue of whether one of the evaluations of Casimir's case should be preferred over the other. The issue hinges on whether the reasons that TR_2 and TR_3 speak of should be direct reasons only, or whether they may also include indirect reasons. Moran doesn't address this issue, and our three Transparency Requirements don't speak to it either. Given the cogency of the distinction between direct and indirect reasons, we may distinguish two versions of TR_2 and two versions of TR_3 – one version of each restricts reasons to direct reasons, the other version of each allows for both direct and indirect reasons. We won't argue that one pair of versions is more in the "spirit" of Transparency than the other pair. We only note what we have already seen, namely that if Transparency requires reference to direct reasons only (let us call that restricted Transparency), beliefs based on forgotten evidence are not Transparent, whereas if Transparency allows for reference to both direct and indirect reasons (let us call that unrestricted Transparency), such beliefs are Transparent.

We thus conclude that if Transparency is to hold for cases of forgotten evidence, either TR_1 must be considered the best formulation of it, or else unrestricted TR_2 or unrestricted TR_3.

With respect to Agnes's testimony-based belief, things are in some respects the same. Her belief is Transparent if that requires the conformity to TR_1, for she can honestly answer "Do I believe Fermat's Last Theorem?" by answering "Is Fermat's Last Theorem true?" Her honest answer to the latter question is likely "yes" and so her answer to the former "I do."

Her belief is not Transparent, however, if Transparency requires the satisfaction of restricted TR_2, for Agnes has no direct reason in favor of the proposition that Fermat's Last Theorem is true *itself.* Nor is her belief Transparent if that requires the satisfaction of restricted TR_3 – for again, Agnes has no direct reasons in favor of Fermat's Last Theorem. However, her belief is transparent on both unrestricted TR_2 and unrestricted TR_3, as her belief in Fermat's Last Theorem is based on testimonial reasons, which are, certainly in her case, indirect reasons.[21]

We assumed at the outset that Agnes' belief is neither alienated nor irrational. This entails that Agnes' belief is a counterexample to Transparency, if that requires for its satisfaction restricted TR_2 or restricted TR_3.

We conclude that beliefs based on forgotten evidence and beliefs based on testimony are Transparent if that requires either the satisfaction of TR_1, unrestricted TR_2 or unrestricted TR_3.

Having now discussed a number of beliefs that stand paradigm for large classes of belief and examined how they fare with respect to each Transparency Requirement; it is time to take stock and reflect on the significance of the results.

5. Conclusion: the significance of the results

We first summarize the results of our discussion so far, and next reflect on their significance.

The results are the following. First, there are counterexamples to Transparency if that requires the satisfaction of TR_3. Beliefs based on non-justifying reasons, long-standing beliefs, basic beliefs (on the assumption of foundationalism), belief in anti-skeptical propositions (on the assumption of internalism) are not Transparent if that requires the satisfaction of TR_3. And, belief in propositions the evidence for which one has forgotten, as well as testimony-based beliefs are not

[21] It seems plausible that most if not all testimonial reasons, i.e., reasons whose content is *that it has been testified that something or other is the case*, are indirect reasons.

Transparent if that requires the satisfaction of restricted TR_3. Yet they often are neither alienated, nor irrational. Given Biconditional and Conditional, this strikes against TR_3 and restricted TR_3 being adequate formulations of Transparency.

Second, there are counterexamples to Transparency if that requires the satisfaction of TR_2. Long-standing beliefs and basic beliefs aren't Transparent if that requires the satisfaction of TR_2; belief in propositions the evidence for which one has forgotten, as well as testimony-based beliefs aren't Transparent if that requires the satisfaction of restricted TR_2. Yet these are neither alienated nor irrational. Given Biconditional and Conditional, this strikes against TR_2 and restricted TR_2 as being adequate formulations of Transparency.

Finally, *all* the cases that we have discussed are Transparent, if that requires the satisfaction of TR_1 – even obsessive beliefs. As we have shown, however, TR_1 also faces problems. It doesn't say, given Biconditional and Conditional, that obsessive beliefs, such as Jane's belief that she will fail the exam, are alienated and irrational (which, intuitively, it is). Nor does it say that beliefs based on non-justifying reasons, such as Elisabeth's belief that her husband will return home, are irrational (which, intuitively, it is). This isn't a flat-out objection to TR_1, but it is an example of a tension between TR_1 and the deliberative stance (to be discussed next).

The significance of these results is that they reveal a tension between Transparency and the central motivation for Transparency, namely the deliberative stance. The tension is that while the best formulation of Transparency is that it requires the satisfaction of TR_1, one can conform to TR_1 without exemplifying the deliberative stance. At the outset, TR_1 seemed to be in line with the deliberative stance, since one concentrates on *p* rather than on the psychological attitude one has vis-à-vis *p*. However, since TR_1 does not specify how QP is to be answered – a simple "Yes" suffices – TR_1 puts no restrictions on the relation between one's belief and one's reasons. Consequently, TR_1 seems to be ill equipped to give a correct verdict on cases of alienated belief, such as obsessive beliefs. Given the intended relation between the deliberative stance and alienation, TR_1 thus seems to be too minimalistic in what it requires to exemplify the deliberative stance.

Hence, the significance of the results is that they present friends of Transparency with a hard choice between the following options:

> [i] Accept that Transparency requires the satisfaction of either restricted TR_2, or restricted TR_3, both of which exemplify the deliberative stance, and accept that large classes of belief, notably long-standing beliefs, basic beliefs (given foundationalism), anti-skeptical beliefs (given internalism), beliefs the evidence for which one has forgotten, and

testimonial beliefs are not Transparent, and, given Biconditional and Conditional, are alienated and irrational. Or,

[ii] Accept that Transparency requires the satisfaction of either unrestricted TR_2, or unrestricted TR_3, both of which exemplify the deliberative stance, and accept that large classes of belief, notably long-standing beliefs, basic beliefs (given foundationalism), and anti-skeptical beliefs (given internalism), are not Transparent, and, given Biconditional and Conditional, are alienated and irrational. Or,

[iii] Accept that Transparency requires the satisfaction of TR_1, that has a very broad scope, and justify why it is detached from the deliberative stance. Or,

[iv] Replace the Biconditional and/or the Conditional with others that, given either TR_2 or TR_3, don't lead to the unwelcome results to which options [i] and [ii] lead.

Option [iii] is for two reasons unacceptable for friends of Transparency. First, because Transparency becomes detached from the deliberative stance, and second because then even paradigm cases of alienated beliefs, namely obsessive beliefs, must be qualified as Transparent. This leaves the friends of Transparency with the other options. How to work it out, is the topic for another occasion.

CHAPTER TWO

Making the Two Topics Problem Transparent

Abstract

Transparency accounts of self-knowledge, which are based on the idea that one learns of one's own mind by attending to the world at large, face the Two Topics Problem (TTP): the problem that the apparent basis for self-knowledge, i.e. *p* (including evidence in favor of *p*), doesn't provide a reason to self-ascribe a particular mental attitude regarding *p*. A careful glance at the state of the debate on transparent self-knowledge shows that there is no consensus of what the relation between *p* and *I believe that p* might be, nor what kind of solution respects the commitments of transparency views that actually establish the source of TTP. The main aim of this paper is to make TTP transparent: to provide a grasp of the nature of the different responses to TTP. The responses that I will discuss are: 1) the view that TTP is only apparent; 2) inferential views; 3) judgment views; and 4) metaphysical views. The upshot is an overview of the necessary choices in the debate. Moreover, it will become apparent that TTP is a problem for understanding conscious mentality more generally and not only for transparency accounts of self-knowledge. This realization, as I will suggest, points us towards a way forward.

1. Introduction

Transparency accounts of self-knowledge, which are based on the idea that one learns of one's own mind by attending to the world at large, face a problem that to many appears unsolvable. According to transparency accounts, one can know that one believes that *p* by attending, in a way to be explicated, to *p* itself. However, there is something outright puzzling about such a procedure. On the face of it, the truth of *p* itself doesn't say much about the truth of the self-ascription *I believe that p*. These

two topics, *p* and *I believe that p*, neither stand in a relation of implication nor of evidential support. The fact that it is raining doesn't entail that I believe that it is raining, nor provides evidence for it: one can imagine numerous scenarios in which it is raining but I do not believe that it is, or in which I believe that it is raining but it isn't. Hence, transparency accounts face the Two Topics Problem (TTP): the problem that the apparent basis for self-knowledge, i.e. *p* (including evidence in favor of *p*), doesn't provide a reason to self-ascribe a particular mental attitude regarding *p* (TTP is further explained in section 2).[1] Friends of transparency accounts of self-knowledge thus need to clarify how the two topics, the proposition and the self-ascription, are related to one another.

A careful glance at the state of the debate on transparent self-knowledge shows that TTP remains unsolved. There is no consensus of what the relation between *p* and *I believe that p* might be, nor what kind of solution respects the commitments of transparency views that actually establish the source of TTP. Discussing the pitfalls and merits of each response proves to be difficult, because the responses involve differences in epistemological, metaphysical, and moral psychological views. The main aim of this paper is thus to make TTP transparent: to provide a grasp of the nature of the different responses to TTP. By discussing the different responses to TTP I hope to provide an overview of the necessary choices in the debate. Moreover, it will become apparent that TTP is a problem for understanding conscious mentality more generally and not only for transparency accounts of self-knowledge. This realization, as I will suggest, points us towards a way forward.

The responses can be divided into four subgroups[2]: 1) First of all, in the literature TTP is sometimes dismissed as a philosophical invention (cf. Cassam 2014, Jongepier 2017). This response presses the question what it is that makes TTP

[1] I will focus on the problem regarding belief and not regarding other mental attitudes. Belief is the most central, and sometimes regarded as the only, case in the literature on transparency. The problem at hand is also known as the puzzle of transparency, the problem of two subject matters, and the evidentialist objection. See, for instance, Barnett (2015), Byrne (2005), Gallois (1996), Martin (1998), Moran (2001), O'Brien (2003), Roessler (2013a), among others. Since its recognition, a number of transparency accounts are proposed specifically in response to this problem.

[2] In general, solutions to TTP have an epistemic and metaphysical dimension: Epistemically speaking, one might think that some transition takes place between thinking *p* and thinking *I believe that p* or one might hold that in certain situations one thought is contained by the other. Metaphysically speaking, one might hold that being in a mental state and knowing one is in that state are distinct mental states or are one and the same state. The result is a conceptual map of four categories of possible solutions: i) transition accounts involving distinct states (portrayed in inferential views and epistemic judgment views), ii) transition accounts involving a single state (this will be exemplified in Boyle's reflectivism in section 5.2), 3) non-transition accounts involving distinct states (an implausible view), and 4) non-transition accounts involving a single state (represented in metaphysical judgment views and constitutive views).

a genuine problem. 2) The second kind of response insists that *p can* be a reason to self-ascribe the belief that *p*. Byrne (2005; 2011), for instance, defends this route, focusing on the epistemological connection between the two topics. 3) Others claim that transparency should be understood as a transition commencing not with the topic *p* but with the *judgment that p*. This subgroup includes accounts by C. Peacocke (1998), A. Peacocke (2017), Roessler (2013a), and Silins (2012). 4) The last kind of response holds that the relation between *p* and the self-ascription should be understood as a metaphysical connection rather than as an epistemic relation (as the first response aims to do). This is the line that Moran (2001) and Boyle (2011; 2019) take.

The paper proceeds in the following way. In section 2, I will discuss the first response to TTP and delineate what makes TTP a genuine problem. Next, I will review and raise problems for the other responses: 2) *inferential views* (section 3); 3) *judgment views* (section 4); and 4) *metaphysical views* (section 5).[3] In the last section, I will present the choices and sketch a way forward (section 6).

2. The roots of TTP

First of all, we should address the question why we couldn't just simply dismiss TTP as a problem by abandoning one of the starting points of transparency. Not everyone is convinced that TTP poses a genuine problem. In the literature TTP is sometimes dismissed as a philosophical invention (cf. Cassam 2014, Jongepier 2017). The idea is that TTP is merely an epistemological problem that is the result of too much rather than too little philosophical theorizing. But if we look more closely at the roots of TTP, we see that they are rational. Appreciating this makes it more difficult to discard TTP as a philosophical invention.

Barnett sees TTP as an epistemological problem of transparency accounts of self-knowledge – a problem he dubs, suitably for his purposes, the evidentialist objection:

> To a first approximation, we usually think that p is a good reason for you to believe that q only if p amounts to strong *evidence* that q is true. This conception of epistemic reasons does not sit well with

[3] Unfortunately, I cannot do as much justice to all accounts as they deserve. Since this paper seeks to provide a broad overview, sometimes I have had to trade depth of coverage for breadth. I hope the tradeoff works, and I hope that the conciseness of the discussion of different views doesn't negatively influence portraying them correctly and sympathetically.

> Transparency, because p often is not very strong evidence that you believe that p.[4] (Barnett 2015, 2)

I agree that putting the problem in epistemic terms adequately captures the puzzling aspect of transparency (hence my own formulation of TTP in the introduction). In transparency accounts of self-knowledge, the challenge is to specify how attending to *p* (and reasons pertaining to *p*) renders the self-ascription *I believe that p* epistemically justified. This is the core question of TTP from a purely epistemological point of view. Seen from such an epistemological viewpoint, why not abandon transparency? From such a viewpoint, the easiest solution to TTP seems to give up the idea that *p* (or reasons pertaining to *p*) are the grounds for self-ascribing the belief that *p*. Hence, on this portrayal of TTP, why not abandon transparency views of self-knowledge?

On this line of thought, however, the gist of TTP no longer relates to the motivations for transparency views of self-knowledge. Given the motivations for transparency views of self-knowledge, the source of TTP is not merely epistemic but is to be found in the first-person perspective and its concomitant rational demands. Moran discusses the problem of TTP in light of 'two quite different types of commitment involved in my avowing a belief of mine' (2001, 74). According to Moran, avowing my belief involves, first, a *commitment of endorsement*: it commits me to the truth of the belief itself – to the world being as depicted by my belief. But, secondly, I also make *an empirical claim* about myself, namely that I believe it. The resulting picture of a self-ascription of belief is that it *is* a report of one's belief, just not *merely* a report. Over and above being a report, a self-ascription involves the commitment to the truth of *p*.[5]

And here we see the non-epistemic roots of TTP, for the commitment of endorsement and the making of an empirical claim put different types of rational demand on self-ascribing a belief. On the one hand, making an empirical claim that I have the belief that *p* requires a relevant epistemic basis: I must secure some epistemic basis to report on my belief that *p*. Only if I have such an epistemic basis for my self-knowledge will my possession of self-*knowledge* be justified. On the other hand, the commitment to the truth of my belief that *p* requires endorsing *p* as

---

[4] Note that Barnett formulates Transparency also in terms of reasons, i.e. only in epistemological terms: '*p* is a good reason for you to believe that you believe that *p*' (2015, 2). Byrne (2005, 95) also formulates TTP merely in epistemological terms. See Martin (1998) and Roessler (2013a) for a different approach.

[5] Cf. Boyle (2011a, 17; 2011b, 231); Moran (2001, especially Ch. 3 and 4); Peacocke (1998, 86). The view that self-ascription of belief commits oneself to the truth of belief is supported by, for instance, Moore's paradox. As many have noted in the debate on transparency, including the previously mentioned authors, transparency and Moore's paradox seem to be two sides of the same coin.

true: I must attend to *p* itself and consider reasons in favor of or against *p*. How can these two requirements, i.e., requirement of endorsement and requirement of securing a relevant epistemic basis, be fulfilled at the same time? Are they even compatible? To comply with the requirement of endorsement, the subject's belief expresses her view of the world and her evidence for taking the world to be that way. If that requirement is to transfer to the self-ascription, then we must somehow explain how that same evidence grounds the self-ascription: i.e., we are confronted with TTP.[6]

This apparent incompatibility between the two requirements for the rational self-ascription of belief forms the basis for TTP. If it is accepted that a self-ascription involves both requirements, then TTP is a genuine problem. Dismissing TTP as some sort of philosophical invention implies dismissing either the requirement of endorsement or the requirement of securing a relevant epistemic basis. Hence, we are confronted with the first choice: either accept TTP as a genuine problem or dismiss one of the requirements.

3. Inferentialist views

The second line of response to TTP is to say that, despite appearances, *p* can be a reason to self-ascribe the belief that *p*. This route is most fervently defended by Byrne, who claims that transparent self-knowledge is acquired by drawing a special kind of inference, namely 'an inference from world to mind' (2011, 203). Following Gallois (1996), Byrne (2011, 204) proposes that transparent self-knowledge of belief is acquired by reasoning in accord with what he dubs the *doxastic schema*:

p
———————
I believe that *p*

Both Gallois and Byrne recognize the abnormality of the schema. 'Plainly the doxastic schema is,' as Byrne writes, 'neither deductively valid nor inductively strong' (2011, 204). On the other hand, Byrne claims, if I follow the schema, then, exceptional cases excluded, the conclusion will be true. If a subject infers that she believes that *p* from the premise that *p*, then her conclusion will be true, because, says Byrne, 'inference from a premise entails belief in that premise' (2011, 206). For

[6] For this portrayal of TTP, see especially Martin (1998, 110-1); Roessler (2013a, 8-10).

this reason, Byrne calls the schema *self-verifying*: reasoning from the antecedent of the schema makes true the consequent. For Byrne, the self-verifying nature of the inference, together with the idea that the beliefs produced by the schema are safe (cannot easily be false), makes the reasoning epistemically justified. And if the schema is self-verifying and if self-verification suffices for epistemic justification, then it seems that Byrne has formulated a solution to the two topics problem.

In evaluating Byrne's account, I want to focus on the crucial assumption that inference from a premise entails belief in that premise. Before discussing this assumption, however, let me first outline various objections to Byrne's account that have been raised in the literature. Most of these objections target the inferential nature of the doxastic schema.[7] Despite the claim that the inference is self-verifying, one might find it problematic that the inference does not fit any standard form of good inference. As Byrne himself notes, it isn't based on deduction, induction or abduction. But that isn't the only way in which the doxastic schema fails to meet conditions that seem to apply to inferences in general. Note, first of all, that the conclusion of the schema can be true regardless of the truth of the premise: if *p* is false but the subject thinks it is true, then the conclusion that she believes that *p* remains true. Even if this seems how it should be in the case of knowledge of one's own beliefs – after all, we have beliefs, and we know we have these beliefs, that are (in fact, and unbeknownst to us) false – this nonetheless constitutes a difference between the doxastic schema and other forms of inference, which are truth-preserving.

Next, it seems to be a necessary condition of rules of inference that they are also valid when used in hypothetical reasoning (cf. Barnett 2015, 12-4; Valaris 2011, 322-3). However, if I merely suppose that *p*, the conclusion that *I believe that p* is false.

Another principle of deductive inference is that your degree of confidence in the premise influences your degree of confidence in the conclusion.[8] The doxastic schema fails to comply with this too: if I am pretty sure about *p*, then I cannot be pretty sure that *I believe that p*. Rather, I probably know that I don't believe that *p*, since I don't take myself to have conclusive evidence in favor of *p*.

Although I think that these objections put severe pressure on Byrne's account and on his claim that the doxastic schema can be understood as an inference, these objections still leave Byrne a way out. After all, Byrne claims that the doxastic schema isn't a *normal* but a *special* inference, which is justified because it is self-

[7] Cf. Barnett (2015), Boyle (2011b), Gertler (2011), Silins (2012), Valaris (2011).

[8] My depiction of the problem follows Barnett (2015, 16-7). For alternative ways of developing the objection, see e.g. Gertler (2011) and Silins (2012).

verifying. Therefore, what "normally" holds for inference need not apply to the doxastic schema.[9] What is needed, then, is an objection where Byrne cannot use the "specialness" of the inference as a way out.

One such objection is Boyle's criticism that only a madman could follow the doxastic schema (2011b, 230-1). Normally, making an inference requires a subject to recognize an intelligible relation between the premise and the conclusion (cf. Boghossian 2014, 4-5; Broome 2013, 228). But although *p* is a reason *to* believe that *p*, it isn't a reason for concluding *that* you believe that *p* (which is why transparency accounts face TTP). Hence, even if the doxastic schema is self-verifying, that doesn't change the fact that the premise isn't intelligibly related to the conclusion, i.e., it implies nor indicates the conclusion. And without an intelligible inferential connection between *p* and *I believe that p*, a subject of sound mind cannot knowingly make the inference. Let's call this *Boyle's madman objection*.

Another possible objection where Byrne cannot use the specialness of the schema as a rejoinder is what I will call the *assumption objection*: it targets the aforementioned assumption underlying Byrne's account. Byrne assumes, in his explanation why the doxastic schema is self-verifying, that 'inference from a premise entails belief in that premise' (2011, 206). This assumption isn't unconditionally true. We quite often reason hypothetically or merely check the validity of an argument: for instance, when we read an inference off the paper in front of us to see whether the conclusion follows from the premises without having any attitudes towards the premises; or listen to someone explaining why they believe something, trying to follow through their reasoning, without having determined for ourselves whether to believe the premises. As Wright has stated, we should 'distinguish inference in general from *coming to a conclusion*…; no particular attitude to [a] proposition is implicit in inference itself' (2014, 28).[10] If this is true, this would rebut Byrne's claim that inference from a premise entails belief in that premise.

This would have implications for Byrne's account. He would no longer be in a position to claim that the doxastic schema is unconditionally self-verifying. Instead, it would only be self-verifying *if* the subject accepts the premise. This means that the subject should only follow the doxastic schema if she accepts the premise. But then Byrne's account seems to presuppose what it wants to explain: knowing that one accepts that *p* is, if not completely identical with knowing that one believes that *p*, at least part of knowing that one believes that *p*. Since this (partial) self-knowledge of one's belief that *p* is presupposed by the doxastic schema, it cannot be the result

[9] For such a response, cf. Setiya (2011, 185ff).

[10] For an argument in support of the view that reasoning doesn't necessarily involve belief in premises and conclusions, see Chapter 3.

of reasoning.[11] This undermines Byrne's project of developing an inferential account of self-knowledge.[12]

One might think that a possible response to both these objections is to go *reliabilist*, where we can take reliabilism as minimally comprising that the subject need not be aware or be able to become aware of whatever it is that justifies her belief in order for the belief to be justified.[13] If the subject need not be aware or be able to become aware of whatever it is that justifies her belief, she doesn't *need* to recognize an intelligible relation between premise and conclusion. Furthermore, on a reliabilist account, one could claim that the subject doesn't need to know whether she accepts the premise as long as she only uses the schema reliably, i.e., only if she actually believes that *p*.[14] No self-knowledge would then be presupposed. Hence, it seems as if a reliabilist take on Byrne's inferential account rebuts both Boyle's madman objection and my assumption objection.

But although reliabilism might work for other kinds of knowledge and for other accounts of self-knowledge, it has problematic consequences for inferential transparency accounts of self-knowledge. In such a reliabilist account, it is not only the case that the subject doesn't *need* to be aware or become aware of the epistemic basis of her belief, but also that she *couldn't* become aware of it. If she would, she would again not be able to recognize an intelligible relation between the premise (the presupposed epistemic basis) and the conclusion. Actually, the subject couldn't even be aware of following a procedure, for if she would, she would need to be able to recognize when following the procedure is appropriate. But since it is only appropriate to follow the procedure if she accepts the premise, this again would presuppose self-knowledge rather than explain it. Thus, such a procedure must, as

---

[11] This also holds for Byrne's rule-following option: if *p*, then believe that you believe that *p*. Supposedly, to follow this rule, one has to *recognize* that *p*, not merely *suppose* that *p*. The question is how it is possible to *recognize* that *p* without presupposing some awareness of *believing* that *p*.

[12] One might think that Byrne could claim that the doxastic schema holds only when the premise *p* is true. This would exclude the possibility of reasoning hypothetically and safeguards the idea that the schema is always self-verifying. However, it would also imply that one cannot know any of one's beliefs that are false. This isn't, I presume, a concession Byrne would be willing to make. A broader implication of this argument concerns inferential accounts of self-knowledge more broadly construed (cf. Cassam 2014; Lawlor 2009). If every piece of reasoning could be a piece of hypothetical reasoning, it seems that inferential accounts would always need to depend on a basis of self-knowledge. In what other way could we know that we *believe* the conclusion of our reasoning? See Concluding Reflections for an exposition of this argument.

[13] Cf. Goldman (1967).

[14] Setiya's non-inferential rule-following account (cf. 2011, 183-6) might be read along these lines. Setiya actually claims that his account isn't reliabilist but see Ometto (2016) for an argument in favor of a reliabilist interpretation. Another reliabilist account, related but different from the kind of account under discussion, can be found in Fernández (2013). For discussion of Fernández' account, see Ashwell (2013b); Coliva (2014; 2016).

Ometto writes, 'always be applied, as it were, behind the subject's back' (2016, 89-90). As a consequence, the subject's self-ascription of her belief that *p* would lack intelligibility from the subject's point of view.[15]

Hence, we arrive at the following evaluation of Byrne's position. Byrne's internalist (non-reliabilist) proposal faces Boyle's madman objection and the assumption objection. If he were to hold on to this account, he would need to accept that the inference is both crazy and lacks epistemic justification, at least without presupposing awareness of one's belief regarding the premise. This seems a very problematic option. However, the other option doesn't fare much better. If Byrne were to transform his account into a reliabilist account, he would need to accept that the procedure can only be applied behind the subject's back. This would imply giving up transparency: the subject's attention to *p* doesn't play a role in achieving self-knowledge. Moreover, such a reliabilist account doesn't provide a solution to TTP, but rather dissolves the problem altogether. If the subject doesn't consciously make a transition from *p* to *I believe that p*, there is no explanation needed of what would make such a transition intelligible for the subject. This implies dismissing the rational demands of self-ascribing the belief that *p*. Hence, the second choice that confronts us is either dismissing Byrne's inferential account or accept the following disjunction, that is, that the inferential procedure is crazy and unjustified (on an internalist construal) or applied behind the subject's back (on a reliabilist construal).

4. Judgment views

The third line of response to TTP is to claim that understanding transparency as a transition from *p* to *I believe that p* is ill-conceived from the beginning. Rather, transparency commences with judging that *p*. If one's judgment that *p* is the basis of (or the same as) one's self-ascription of belief, so the thought goes, there is no transition between the two original topics. This means that TTP wouldn't even surface as a problem at all. So, initially, judgment views seem to have good prospects in solving TTP.

[15] This also points to a deeper problem concerning the methodological assumptions underlying such an approach. As Roessler (2013a, 13-4) writes: 'If we adopt a radically externalist approach, we should not expect to be able to discover the basis of second-order beliefs simply through reflection on what we intentionally do when we reflect on our own beliefs, any more than we should expect to discover the non-conscious mechanisms underpinning vision by intently looking at the world.'

There are several judgment views of transparent self-knowledge out there that harbor many useful and intricate insights. Given the goal of this paper, however, there is no need nor space to discuss them extensively. The reason for this is the following: no matter the posited connection between judgment and belief, all judgment views hold that self-knowledge of judgment gets us to self-knowledge of belief.[16] This means that these views rely on the assumptions that 1) we have self-knowledge of judgment and 2) self-knowledge of judgment can be accounted for in a way distinctive from self-knowledge of belief. It is these assumptions that need to be addressed in this section.[17]

To start, what motivates the claim that self-knowledge of judgment and self-knowledge of belief are distinct? After all, both judging that *p* and believing that *p* are characterized by *assenting to* p. If judging and belief are to be different, then, they must be different in the kind of role they play in our mental economy. Hence, a basic assumption of judgment views is that judgment and belief are different kind of mental items: judging is a mental (and conscious) act, whereas believing that *p* is a standing attitude, it is not something we do (viz., a mental act) or undergo (viz., a sensation) (cf. C. Peacocke 1998, 88). Judgment views need to rely on this assumption so that self-knowledge of judgment and belief can come apart.

However, the distinction itself doesn't yet explain how we know our judgments. We find two different kind of accounts in judgment views: the first focuses on the idea that, because judging is a mental act, it is conscious, and therefore there is something it is like to judge. In other words, the first option seeks to account for our self-knowledge of judging on the basis of its phenomenal qualities. The second option zooms in on the idea that judging is an *intentional* mental act, and

[16] There are two broad categories of judgment views: *epistemic* judgment views and *metaphysical* judgment views. Epistemic judgment views maintain that my judgment that *p* forms an epistemic basis for my self-ascription of my belief that *p*. Judging that *p* is thus a reason for self-ascribing the belief that *p*. Some claim that I *infer* that I believe that *p* from the fact that I judge that *p* (cf. Cassam 2014), while others claim that my self-ascription of belief is non-inferentially based on my judgment (cf. C. Peacocke 1998; Silins 2012). Cassam's view will not figure in the discussion of judgment views, because ultimately it is based on an inferential view of self-knowledge. Silins clearly holds a judgment view – 'judgment is a guide to belief' (2012, 297) – but will not figure in the discussion because he assumes that we have phenomenological awareness of judgment without giving an account of how he thinks this works. Metaphysical judgment views, by contrast, claim that my self-ascription isn't based *on* but based *in* my judgment – the judgment manifests my self-knowledge of belief (cf. A. Peacocke 2017; Roessler 2013a).

[17] In the literature on transparent self-knowledge and TTP, there is extensive discussion on the correct metaphysics of judgment and belief (cf. Boyle 2009a; Peacocke 1998; Schwitzgebel 2010; Cassam 2014). According to some, if one holds a view of judgment and belief, where judgment is a conscious act and belief a standing attitude, this would dissolve TTP. What the discussion of judgment views will make clear, however, is that one's metaphysical presuppositions do not actually dissolve TTP. Even if all transparency theorists accept the claim that judgment is a conscious act (and different from belief), they still face the challenge of explaining how one is aware of one's judgment.

that we thus have *action awareness* of judging. I will discuss two possible ways to spell out action awareness in more detail, one based on *contrastive awareness* and one based on *proximal intention*. I will discuss these three options in section 4.1, 4.2, 4.3, respectively.

One last general point concerns the possibility that an account of self-knowledge of judging might also face a problem analogous to TTP. Advocates of judgment views maintain that the content of the judgment also plays an important role (cf. C. Peacocke 1998, 74, 87; 2009, 211-2; A. Peacocke 2017; Roessler 2013a). But, intuitively, if the content of the judgment that *p* is to play a role, then one is to attend to *p* in acquiring self-knowledge of that judgment. But no matter how closely a person attends to *p* (i.e., the proposition or world represented in the proposition itself), she will not find any evidence pertaining to the state of her mind. Parallel to the case of belief, the truth of *p* doesn't seem to be a good reason for a person to self-ascribe that she *judges* that *p*. On this line of thought, then, explicating self-knowledge of judging requires a solution to a problem analogous to TTP. Hence, the task for judgment views is to give an account of self-knowledge of judgment, where the content of the judgment plays a role, but where that doesn't result in (a problem analogous to) TTP.

4.1 Phenomenal quality

On the first view, one is aware of one's judgment that *p* because of its phenomenal quality: there is something it is like to hold the content "*p*" in one's mind while taking *p* to be true. One proposal by Christopher Peacocke is that one is aware that one judges that *p*, because of one's view of *p*. And one is aware of one's view that *p* because of its distinctive phenomenal qualities (cf. C. Peacocke 2007). Such an account wouldn't face TTP, because the self-ascription isn't based on *p* but on the phenomenal qualities of one's view that *p*. Being aware of the phenomenal qualities of *p* is a form of what I call *attitudinal awareness*, i.e., awareness of the attitudinal aspects of one's mental attitudes instead of an awareness of (one's commitments regarding) the content of one's attitudes.

One basic worry here is that the phenomenology of our propositional attitudes isn't sufficiently fine-grained to set judging that *p* apart from other propositional attitudes, such as supposing that *p* or wishfully thinking that *p*. I don't know of a way to establish this negative claim, but neither do I think that C. Peacocke is in a position to establish the positive claim. As Maja Spener (2011) argues, although the orthodox assumption that propositional attitudes lack phenomenal quality has lost its dominance, the idea that cognitive phenomenology is sufficiently "thick" or distinctive to make subtle distinctions between different kinds of

propositional attitudes remains highly controversial.[18] Moreover, following Spener's argument, the disagreement underlying this controversy has consequences for establishing such a negative or positive claim, because the only means available to establish such a claim (the qualitative experience of propositional attitudes) is precisely what the disagreement is about: for example, C. Peacocke introspectively judges that his "judging" comes with a distinctive phenomenal quality, whereas, for instance, Coliva (2016, 103ff) judges that her "judging" doesn't. Since there isn't any independent way of accessing such cognitive phenomenal qualities, the fact that people disagree about the right characterization of it, makes their introspectively formed judgments suspect and at least lowers their credibility (cf. Spener 2011, 280-2). So, although I don't know how to establish that C. Peacocke's claim is false, neither do I think that he is in a position to make the positive claim, i.e., that our cognitive phenomenology is sufficiently fine-grained to set judging that *p* apart from other propositional attitudes.

Another problem concerns the compatibility of C. Peacocke's proposal and the requirement of endorsement. In his account, one's view of *p* ultimately doesn't play a role in the epistemic justification of one's self-knowledge of judging that *p*. Instead, one's epistemic basis consists of being aware of certain phenomenological qualities. By introducing a form of attitudinal awareness, a person's actual view of *p* isn't bestowed any epistemic status. This difficulty receives more attention in the next section(s).[19]

4.2 Contrastive awareness

The second proposal (this section) and third proposal (next section) focus on the idea that if judging is a mental act, then we have action awareness of judging. A person's self-knowledge of judging would then consist in action awareness, and not be based (merely) on *p* (cf. C. Peacocke 2009; A. Peacocke 2017; Roessler 2013a). This proposal only works if judgment is taken to be an *intentional* mental action, because one can have action awareness only of those actions that are intentional.

The basic problem with thinking of judging as an intentional mental action, however, is that it is obscure what kind of intention could be involved in judgment. I can intend, for example, to answer a particular question, but I cannot have the intention *to make a particular judgment*. One cannot intend to judge that *p* without

[18] Cf. Spener 2011. For an excellent edited volume on cognitive phenomenology, see Bayne and Montague (2011).

[19] Other objections to C. Peacocke's account have focused on the connection between judging that *p* and believing that *p* (cf. Boyle 2019). As outlined in the introduction, for the sake of the discussion, I only focus on his account of self-knowledge of judgment.

already judging that *p*. Similarly, one cannot intend to think a particular thought without already thinking it (cf. Doyle 2018; C. Peacocke 1999). Hence, if judging is to be an intentional action, there must be something other than having an intention to make a specific judgment that makes it so. The task for judgment views is thus to show in what sense judgments are intentional mental actions and how their sense of being intentional yields self-knowledge of them. I will discuss two proposals, one by A. Peacocke (2017) and one by Roessler (2013a), that seek to explain this.

How does Antonia Peacocke address the issue of self-knowledge of judgment? Her account, in short, is the following. She claims that judging isn't always but *can be* an intentional action: it is possible, for instance, to set out to determine what is true, to determine whether *p* is true (i.e., a specific content) or to make a judgment about a particular topic (A. Peacocke 2017, 362). Based on this idea of the intentionality of judgment, A. Peacocke claims that when judging is intentional, a subject is aware that she makes a judgment. Such awareness, says A. Peacocke, doesn't latch onto the propositional content of the judgment, but onto its attitudinal aspect (cf. A. Peacocke 2017, 362). She calls this *contrastive awareness*, meaning one is aware of judging *rather than* being engaged in some other mental activity: 'you can be aware, that is, that you are doing *this* sort of thing (e.g. imagining one's wedding) rather than *that* sort of thing (e.g. recalling one's wedding)' (ibid.).[20] Contrastive awareness also is a form of attitudinal awareness.

There are several problems with this proposal. First of all, *normal* intentional action seems to be of the wrong *kind* to include judging. Normal intentional action is action with temporal extension – one that progresses in time. This doesn't seem to apply to judgment. As Roessler points out:

> Answering a question may take time. It may, as Evans emphasized, involve many kinds of mental activity, such as observation, deliberation or recollection. You may be in the process of answering the question whether *p* without ever reaching a verdict, due to being interrupted. On the other hand, it is hard to think of what might be involved in being interrupted in judging that *p*. (Roessler 2013a, 3)

Most intentional actions, such as the one mentioned by A. Peacocke (e.g., determining whether *p* is true), may be actions that take time. By contrast, the act of judgment marks the endpoint of these former actions and does not itself take any

[20] C. Peacocke (2007; 2009) also seems to have something like this in mind. As he writes (2007, 365): 'it is a feature of your consciousness that you are, for instance, judging something rather than forming an intention.'

time.[21] If judging isn't a kind of activity that takes time, this puts doubt on whether contrastive awareness is the right kind of awareness.[22]

A second problem for A. Peacocke's proposal concerns the idea of contrastive awareness itself. Being contrastively aware of judging means being aware of *judging* rather than, for instance, *hypothesizing*. The question is, however, how *this* kind of contrastive awareness is related to a contrast in content, e.g., knowing whether one judges that *p* rather than not-*p* or *q*? How is the content of the judgment related to contrastive awareness? A. Peacocke motivates the idea of contrastive awareness by referring to uncertainty about one's judgment:

> If, in your passive train of thought, you did in fact come to a judgment that your daughter has no real skill at tennis, you might simply not know whether that was a real judgment or merely a case of entertaining a hypothetical. Crucially, the uncertainty in either case would attach to the attitudinal aspect of the thought… (2017, 362)

According to A. Peacocke, then, the uncertainty pertains to the attitudinal aspect of the mental action, and, conversely, any certainty about what you're doing too. A. Peacocke here seems to presuppose that you know a particular *thought* but that you remain uncertain about whether that thought was a judgment or an instance of entertaining a hypothetical. However, the examples she gives of judgment as

---

[21] The temporal aspects of judgment are also discussed in, for instance, Geach (1957, 104); Soteriou (2009, 240ff).

[22] One might object that there are non-mental actions that do not seem to take any time, but of which we can be aware nonetheless. As C. Peacocke (2009, fn. 9) writes: 'One can have awareness of something that does not take time, both in the bodily and in the mental domains. Stopping talking can be an action, and the agent can have an action awareness of it. It is not a continuing event. Judging and deciding are also not temporally extended processes, but the subject can have an action awareness of them too.' I think this underestimates the distinctive nature of judgments. Judgments are such that they could not have any duration, whereas stopping talking *can be* something that takes some time. For instance, the following sequence might be involved in the action "stopping talking": one might hesitate in making a last reply, utter half a word, and then keep silent. Now, the real question seems to be what constitutes the difference in temporal possibilities between judgment and stopping talking? It isn't just that judging isn't a process in time, but also that it doesn't seem to be a *production* of something at all (whereas "stopping talking" might be said to produce silence, closed lips, makes one's vocal cords come to rest, or a change in the conversation). What could judging produce? Perhaps one thinks that an act of judgment produces the thought that *p*, where *p* is represented as true. But what would an act of judgment be if not the thinking of *p* as true? Distinguishing the two seems to make an act of judgment like a magical trick of the mind. If judging doesn't produce anything, it's also difficult to see how it can involve an intention *parallel* to normal intentional action. Compare Soteriou (2009, 244): 'the mental act of judging does not seem to require the production of anything. This is why although it makes sense to ascribe to the agent an intention to assert that p [for it produces spoken word], it doesn't make sense to attribute to the agent an intention to judge that p.' For more on the *unproductive* character of judgment (and belief), see Boyle (2009a, 32ff).

intentional, e.g., to determine whether *p* is true, are cases where it is clear that one sets out to judge, but unclear what the actual judgment will be.

To clarify my point, consider the way in which A. Peacocke relates a person's contrastive awareness to Anscombe's question 'Why?'. She writes that 'if you are intentionally assessing your daughter's skill in tennis, I can ask you why, and you can tell me that you need to decide which lessons to book for her' (2017, 362). The question "Why?", as introduced here, addresses the reason for being engaged in the activity of judging. It thus relates to the point of "determining whether *p* is true." But, as such, it doesn't bring out anything about your actual views on *p*. For it doesn't address the reason for judging that *p*, i.e., the reason in favor of *p*. Again, this might put doubt on the adequacy of contrastive awareness, especially in relation to transparency. If contrastive awareness isn't related to one's view on the matter, then it is unclear how one's attention to *p* plays a role. Again, the requirement of endorsement is at stake.

4.3 Proximal intention

Another proposal to explain action awareness of judgment has been put forth by Johannes Roessler (2013a). He claims that being aware of judging is being aware of the intention inherent in judging, namely 'to express one's conviction that *p*' (Roessler 2013a, 3). As we have seen, this cannot function as a *normal* (prior/prospective) intention, because that would presuppose knowledge of one's belief that *p* (knowing one's conviction that *p*). But according to Roessler, we shouldn't think of the intention inherent in judging as a prior intention. The intention doesn't have to preexist before the act of judging. Following Williams' remarks on assertion, Roessler maintains that a judgment 'can be spontaneous "as to what," even if it is not spontaneous "as to whether"' (2013a, 4). He gives the following example:

> Compare the case of trying to recall Hume's date of birth: you may no sooner acquire the intention to say '1711' than you blurt it out… Your saying '1711'…may be premeditated 'as to whether' – it may be the realization of a prior intention to say when Hume was born – yet spontaneous 'as to what': you may not have a prior but only a *proximal* intention to say '1711'. (Ibid. My italics.)

Thus, according to Roessler, judging is intentional in virtue of the *proximal* intention to express one's conviction that *p*. This leads him to claim that when one makes a judgment, one is aware of this intention, and since this intention expresses one's

belief, one will thus be aware of one's belief. In a slogan: judging is 'a matter of intentionally (hence knowingly) expressing one's belief that *p*' (Ibid.). Supposedly, this doesn't presuppose self-knowledge of one's belief, because the intention inherent in judgment is a proximal intention.

However, we aren't told what such a proximal intention actually is. In general, "proximal" might indicate that the intention didn't exist long before its execution (the intention is nearby the action), or it might indicate that the intention and action occur simultaneously. Now, it seems that Roessler can only have the last connotation in mind – otherwise the intention would still presuppose self-knowledge. However, calling an intention proximal (simultaneous) is not yet to explain why the intention *couldn't* exist at all prior to its execution. That is to say, the intention to express one's conviction that *p* not only *happens* to occur simultaneously with the act of judgment, but it *cannot* occur prior to the act of judgment. For if it would, it would presuppose self-knowledge of belief. But this makes it an odd kind of intention, because, normally, the functional role of an intention is such that the intention to do something in the future transfers into the intention with which one is doing something. Additionally, normally an intention seems to bring out the point of an action: it is an answer to Anscombe's famous question "Why?" But if we ask why you judge that *p*, it doesn't seem to make sense to say "in order to express my conviction that *p*." Rather, the relation seems to be the other way around: if we ask why you express your conviction that *p*, you might say "because I judge *p* to be true." On this line of reasoning, to intend to express one's conviction that *p* is made intelligible by believing that *p* rather than the other way around. So it seems that more work needs to be done to explain what a proximal intention is on Roessler's account, what kind of relation it has to the act of judgment, and in what way its function resembles the role of intention in "normal" intentional action.

What I have been arguing, then, is that judgment views seek to solve TTP by categorizing belief as standing attitude and judgment as mental act, and by claiming that, therefore, self-knowledge of belief and judgment come apart. Even if one accepts this distinction, it remains an open question how judgment views account for self-knowledge of judgment. All three proposals – phenomenal quality, contrastive awareness, and proximal intention – leave key points unexplained. With regard to the first proposal, it remains unclear how C. Peacocke could establish the positive claim that our cognitive phenomenology is sufficiently fine-grained to set judging that *p* apart from other propositional attitudes. As to the second proposal, contrastive awareness seems to be the wrong kind of awareness: both to be awareness of judging (for judging isn't temporally extended) and to be awareness of your view of the content of the judgment (and thus be transparent). Finally,

Roessler's proposal requires more work on the nature of proximal intention. The challenge here will be to explicate it in a way that its function resembles the role of intention in "normal" intentional action, so as to avoid any ad hoc adjustments to the notion of intention. Hence, the third choice we face is to either dismiss judgment views of self-knowledge or address any of these problems satisfactorily.

Importantly, this means that the assumption that TTP would dissolve if the transition involved in Transparency would be understood as a transition between judgment and belief is false. Without a satisfying account of self-knowledge of judgment, we are still left where we started. This means that – pace the discussion in the literature on transparent self-knowledge and TTP, where the correct metaphysics of judgment and belief play a central role (cf. Boyle 2009a; Peacocke 1998; Schwitzgebel 2010; Cassam 2014) – one's metaphysical presuppositions do not dissolve TTP. Even if all transparency theorists accept the claim that judgment is a conscious act (and different from belief), they still face the challenge of explaining how one is aware of one's judgment.

5. Metaphysical views

The final kind of solution to TTP maintains that understanding the relation between *p* and *I believe that p* requires a metaphysical approach instead of an epistemological one or one involving the introduction of a difference between self-knowledge of judgment and belief. The central idea is that the proper understanding of the *nature* of what is involved in transparency – e.g., the nature of belief, mental agency, or awareness – will also account for the *epistemic credentials* of transparency. I will not discuss full-blown *constitutive views*, such as Shoemaker's account (2009), because even if such accounts would be compatible with the phenomenon of transparency, Shoemakerian self-knowledge is independent from going through any transparency procedure. Hence, they aren't *transparency* accounts of self-knowledge. Rather, I will focus on Richard Moran's (2001) *deliberative account* in which he invokes the nature of rational agency to solve TTP and on Matthew Boyle's *reflectivism* (2011; 2019), which he recently explicated as involving a specific metaphysics of awareness.

Importantly, both of these transparency accounts hold that mental activity with the content *p* and mental activity with the content *I believe that p* can be instances of a single mental attitude, namely consciously believing that *p*.[23] This also means that there cannot be an inference between two wholly independent facts (or

[23] Cf. Moran (2001, 27-32) and Boyle (2011, 233; 2019).

topics). As Boyle writes, transparent self-knowledge should not be understood as 'knowledge of one realm of facts [that one arrives at] by inference from another, epistemically independent realm of facts' (Boyle 2011, 233). However, that one doesn't infer one fact from the other doesn't imply that self-knowledge doesn't require a cognitive achievement. For Moran, such cognitive work is done by making up one's mind. For Boyle, the cognitive achievement consists in a reflective *step* or *transition* from the world-directed thought to the self-ascription.

5.1 Moran's appeal to agency

Moran's solution to TTP is not easy to pin down. Its starting point is an appeal to rational agency. According to some, this appeal turns Moran's account into an inferentialist account. I will resist this reading and propose another interpretation based on abilities. What remains problematic, on the latter reading, is that it doesn't explain the second requirement, i.e., of securing a relevant epistemic basis for the self-ascription.

Moran's transparency account remains close to Gareth Evans' observation that when we are asked whether we believe that *p*, we answer that question by answering the question whether *p*: 'When asked "Do I believe P?," I can answer this question by consideration of the reasons in favor of P itself' (Moran 2003, 405; cf. 2001, 62ff.). Moran immediately recognizes that such a formulation of transparency raises a problem, namely TTP. Here is a concise description of Moran's view of the problem and its solution.

> What right have I to think that my reflection on the reasons in favor of P (which is one subject-matter) has anything to do with the question of what my actual belief about P is (which is quite a different subject matter)? Without a reply to this challenge, I don't have any right to answer the question that asks what my belief is by reflection on the reasons in favor of an answer concerning the state of the weather. And then my thought at this point is: I would have a right to assume that my reflection on the reasons in favor of rain provided an answer to the question of what my belief about the rain is, if I could assume that what my belief here is was something determined by the conclusion of my reflection on those reasons. An assumption of this sort would provide the right sort of link between the two questions. (Moran 2003, 405)

Moran here seems to introduce a *linking assumption* between the two topics (what he calls subject-matters), namely that my belief regarding *p* (and thus the right

answer to the question whether I believe that *p*) is determined by my answer to the question whether *p*. Since Moran thinks we are entitled to answer the inward-directed question by answering the relevant outward-directed question, this should be an assumption that he is willing to make. Consequently, some[24] interpret Moran as claiming that one acquires self-knowledge of one's belief that *p* on the basis of the following inference: 1) I conclude that *p* and 2) If I conclude that *p*, then I also believe that *p* because it is determined by my conclusion (i.e., the linking assumption), so 3) I believe that *p*. This would turn Moran's account into an *inferentialist judgment account* and would face the same problems as discussed in the previous section. To wit, on this reading, Moran's account would need to presuppose 1) that we have self-knowledge of what we conclude (or judge), and 2) that knowing that I conclude (or judge) that *p* is different from knowing that I believe that *p*.

However, given Moran's explicit endorsement of the epistemic immediacy of self-knowledge (i.e., self-knowledge isn't based on observation or inference) and his explicit aspirations to explain this immediacy, such an interpretation of Moran is untenable. What's more, he doesn't seem to claim that the linking assumption would have to play an explicit role in acquiring self-knowledge. Rather, Moran seems to hold that merely having the relevant abilities entitles one to make the transition from *p* to *I believe that p*. For instance, he writes that 'such entitlement is…a matter of possession of the relevant practical and cognitive abilities' (2003, 412). Here, the relevant ability seems to be the ability to make up one's mind, i.e., that one's beliefs are determined by one's conclusions. A different, and in my mind a better reading of Moran's account thus is: insofar as one is able to make up one's mind (determine one's belief) by reflecting on the reasons for or against *p*, one is entitled to make the relevant self-ascription.

But a solution in terms of having the requisite abilities seems wanting as a solution to TTP. It seems wanting because it leaves us in the dark about how the subject is supposed to grasp the intelligibility of her self-ascription. Why would having this ability suddenly render the self-ascription intelligible to the subject?[25] Moran claims that a self-ascription involves two requirements, i.e., of endorsement and of securing a relevant epistemic basis, but his solution seems to discard the second. Without giving a more substantial account of how the agential abilities to make up one's mind make it intelligible for a subject to make an empirical claim about her mind (to report on her mental state), such a report seems to appear from

[24] For this interpretation of Moran, see, for instance, Finkelstein (2012, 107) and Cassam (2014, 103).

[25] As a reminder, it isn't possible to claim that the subject can *assume* that she has this ability and use that assumption in acquiring self-knowledge, for then self-knowledge is no longer epistemically immediate.

nowhere.[26] And so it seems that the self-ascription can only be intelligible to the subject as a way of expressing the conclusion of her reflections on *p* – that is, as an expression of her conviction that *p*, not as a *report* on that state.

5.2 Boyle's reflectivism

Where in earlier work Boyle defended and further developed Moran's account, his most recent writings suggest a new transparency approach, called *reflectivism*, that puts the nature of awareness central stage.[27] As we will see, Boyle provides a solution to TTP by postulating that the nature of awareness is such as to implicitly comprise self-awareness. I think Boyle's account puts the action precisely where it should be, although more work needs to be done to explain why the resulting account is a transparency account of self-knowledge.

The basic idea of Boyle's account is that the transition in transparency is not an inference from *p* to *I believe that p* but that one can, through a reflective act, come 'to explicit acknowledgement of a *condition* of which one is already tacitly aware' (Boyle 2011, 227).[28] This is possible, Boyle claims, because a world-directed thought encompasses tacit knowledge about the mental attitude. Following Sartre, Boyle distinguishes *what* one thinks about (positional consciousness) from the *way* in which one thinks about it, e.g. believing versus hypothesizing (non-positional consciousness or the mode of presentation) (2019, 17ff). According to Boyle, the information about this mode of presentation is already tacitly present when one focuses on the presentation itself.

---

[26] In the case of intention, more work has been done on what such a substantial account could be. Philosophers such as Hampshire (1975), Stroud (2003), and Roessler (2013b, 47-8) claim that it's a *structural element of deliberation* that one's practical reasoning warrants both the formation of an intention and an empirical statement about the future. But what is this structural element supposed to be? How would it provide a solution to TTP? Without settling these latter questions, saying that it's a structural element of deliberation is just another name for what we seek to explain.

[27] For a defense and development of Moran, see, for instance, Boyle (2011a; 2011b). For his most recent account, see Boyle (2019).

[28] First, this statement was based on the claim that '…in the normal and basic case, believing *P* and knowing oneself to believe *P* are not two cognitive states; they are two aspects of *one* cognitive state – the state, as we might put it, of knowingly believing *P*' (2011b, 228). Recently, he added his account of awareness to this, so I will focus on that. A puzzling aspect of Boyle's proposal is that there are two thoughts, but only a single cognitive state. Boyle doesn't explicitly consider this, but he must take it that the same cognitive condition may find expression in distinct thoughts. That is, being in a condition of *believing that p* can be expressed in a world-directed way – e.g., there will be a third world war – or as self-ascription – e.g., I believe there will be a third world war. The difference between the two is the concepts being used (epistemic), not the mental state that they express (metaphysical). One might be hesitant, however, to give up the plausible idea that two thoughts using different concepts and thus containing different information imply distinct mental attitudes. Namely, one attitude vis-à-vis the proposition *p* and one vis-à-vis the proposition *I believe that p*.

In the case of belief, the idea is as follows.[29] If one believes that *p*, one holds *p* to be true (Boyle 2011, 236). This means that if one has the thought that *p*, *p* is not just a 'neutral' proposition, but somehow presented as true or correct. As Boyle writes:

> Suppose I wonder whether there will be a third world war and reach the alarming conclusion that (5) [t]here will be a third world war. *What* I conclude here is a proposition about the non-mental world, but my manner of representing this proposition differs from the way I would represent it if I were merely supposing (5) for the sake of argument… [S]ubjects who can deliberate competently… must be able to distinguish between a factual question being open and its being closed: between the attitude toward *p* in considering *whether p* and the attitude involved in settling this question one way or another… The point here is not merely that the subject's answer to the question whether *p* expresses a belief she holds, but that she herself already implicitly distinguishes between this mode of presentation and a contrasting non-committal mode…We might therefore say that in concluding that there will be a third world war, she expresses a *non-positional* consciousness of her own belief: an awareness that figures, not as object of her thought, but as the necessary background of her thinking of the question of whether there will be a third world war as settled. (Boyle 2019, 23-24)

That is to say, if a person reaches a settled view on *p*, then the question whether *p* is presented as *resolved* or *closed*. This contrasts with questions that remain *open*. If a question remains unresolved, then what is presented as open is not whether there will be a third world war, whether the fact itself is indeterminate, but rather, what is presented as open is a person's stance regarding the proposition. Similarly, if a question is resolved, then she implicitly represents her belief as determinate. The idea is thus that world-directed thoughts aren't merely world-directed. Boyle postulates that a single thought, in presenting us with an object, necessarily includes

[29] Boyle explains his views by starting with the case of perception. According to Boyle, other people find his views most convincing in the case of perception (personal communication). In the case of perception, the idea is the following: suppose I perceive a purring cat in front of me and have the thought "This cat is purring". In thinking this thought, it is presented to me in a certain way. I could not have the thought "This cat is purring" in the same mode, if *this cat* was not perceived by me. The use of the demonstrable *this* already contains that *I am perceiving* the cat rather than imagining a purring cat or hoping for one (cf. 2019, 21-22).

information about the way in which this object is presented, although this manner of presentation is not (yet) an object of consciousness itself (it remains tacit).

There is something very intuitive about this proposal. It indeed seems as if we shouldn't strictly separate the thought that *p* and the way in which *p* is presented. But what remains unclear is how the tacit information about one's mental state becomes explicit. Boyle (2011) argues that it becomes explicit through a reflective act: a reflective judgment based on the way the proposition is presented. The reflective act consists of a transition from *believing that p* to *reflectively judging I believe that p*. This reflective step should not be regarded as an inference. Rather, the self-ascription makes explicit certain information that was already contained in the world-directed thought.

The question is whether this reflective step can be understood in a way such that it both provides a solution to TTP and satisfies the intuitions of transparency. The reflective step consists of, as Boyle writes, 'shifting one's attention from the world with which one is engaged to one's engagement with it' (2011, 228). A subject thus shifts her focus from *p* to the way in which *p* is presented. Furthermore, '[w]hat justifies this reflective step will be, not the sheer thought that there will be a third world war, but her non-positional consciousness of her own stance on this question' (Boyle 2019, 24). Now, one might become suspicious here, for notions such as "shifting attention" remind us of a perceptual model of self-knowledge. What is implied in "shifting one's attention"? What does it mean that the reflective step is grounded in the subject's non-positional consciousness?

What seems to be implied is that the basis of self-knowledge isn't the world-directed thought as such, but rather the information about one's mind that is, as postulated by Boyle, inherent in that thought. What's doing the trick, so to say, is not attending to *p*, but paying close attention to the way in which *p* is presented. What happens, for instance, if the subject doubts whether her self-attribution is correct? What would the subject do if she is uncertain about her self-ascription? What a subject should do if the self-ascription is to be transparent, is captured in the following quote by Evans:

> …when the subject wishes to make absolutely sure that his judgment [i.e., that I believe that *p*] is correct, he gazes again *at the world* (thereby producing, or reproducing, an informational state in himself); he does not in any sense gaze at, or concentrate upon, his internal state. His internal state cannot in any sense become an *object* to him. (He is *in* it.) (Evans 1982, 227)

What would Boyle's subject do? Would she, in checking her self-ascription, *gaze again at the world*? If the tacit information about the way in which *p* is presented is the source of self-knowledge, it seems that the subject wouldn't. Instead, she would turn again to the question whether *p* is presented as an open or closed question. However, this question pertains to the mode of presentation and not to *p* itself.

It thus seems that the actual basis of self-knowledge isn't the world-directed thought, but the information about the way in which this thought is presented. This would mean that, in the end, it's being aware of the *mode of presentation* in Boyle's reflectivism that is doing the trick. As such, Boyle seems to introduce an attitudinal form of awareness to solve TTP. And although I am inclined to agree with Boyle that a solution to TTP should focus on the nature of awareness, it remains to be seen in what way the resulting form of awareness still counts as a transparency account of self-knowledge – an account where attending to *p* is crucial in acquiring self-knowledge. Especially, if it is to count as a transparency account, it should be able to explain that any uncertainty in one's self-ascription would direct one's attention back to the content of one's self-ascription (i.e., gazing back at the world), not to the mode of presentation.

To conclude, the two varieties of the metaphysical solution discussed in this section provide quite distinct responses to TTP. Moran hopes to solve the problem by referring to the nature of agential abilities, which supposedly enable us to avow and at the same time report a mental attitude. But that these abilities seem to enable us to do this is not yet an explanation of *how* they enable this. Boyle, by contrast, appeals to the nature of awareness to solve TTP. Awareness of one's belief that *p* involves, according to Boyle, being explicitly aware of *p* itself and implicitly aware of the way in which *p* is presented (as an open or closed question). With this latter attitudinal form of awareness, Boyle seems to introduce a form of non-transparent awareness. Hence, the fourth choice that confronts us is threefold: either we dismiss metaphysical solutions or accept that it is a structural element of reflective abilities that they enable avowal and report or accept that transparent awareness should be supplemented with a form of attitudinal awareness. The latter two options still face us with unsolved questions.

6. Concluding remarks and a possible way forward

The starting point of this paper was that TTP seems to remain an unsolved problem for transparency accounts of self-knowledge. The aim of this paper was to provide

an overview of the different responses to TTP, their pitfalls and merits, and specifically the choices they present us with. These choices are the following:

1) Either accept TTP as a genuine problem or dismiss one of the requirements.
2) Either dismiss Byrne's inferential account or accept the following disjunction, that is, that the inferential procedure is crazy and unjustified (on an internalist construal) or applied behind the subject's back (on a reliabilist construal).
3) Either dismiss judgment views of self-knowledge or address any of their problems satisfactorily.
4) Either dismiss metaphysical responses or accept that it is a structural element of reflective abilities that they enable avowal and report or accept that transparent awareness should be supplemented with a form of attitudinal awareness (and solve the remaining problems of these latter two options).

In my view, choices 1) and 2) are more easily made than choices 3) and 4). First, since I don't see any easy way to dismiss either the requirement of endorsement or the requirement of securing a relevant epistemic basis, I take it that TTP is a genuine problem. Secondly, either option of the disjunction in choice 2) seems to me outright problematic and to bear too many unwanted consequences. But as to choices 3) and 4), matters are less clear. Each proposal involves, in one way or another, the explication or postulation of some form of attitudinal awareness – to wit, phenomenal awareness (C. Peacocke), contrastive awareness (A. Peacocke), awareness that comes with a proximal intention (Roessler), awareness that springs from the structure of deliberation (Moran), and awareness of the mode of presentation (Boyle). That all these accounts need to postulate such a form of awareness seems to be the result of the following key question: how do we go from being aware of the *content* of a mental state to being aware of the mental state itself (i.e. to being aware of *the kind of awareness* itself). This question actually mirrors a question that was raised in discussions of the nature of awareness (and not specifically in discussion of TTP). In fact, it brings us back to so-called *phenomenal transparency* or *diaphanousness* of mental states: i.e., the appearance that what we are aware of is the *content* of our mental states and not the *awareness* itself.[30] As G. E. Moore stated:

[30] Hofmann (2018), Kind (2003), Paul (2014).

> …that which makes the sensation of blue a mental fact seems to escape us: it seems, if I may use a metaphor, to be transparent – we look through it and see nothing but the blue… (Moore 1903, 446)[31]

And how could it be different? What could awareness of awareness consist in; what would that be an awareness *of*?[32] As is forcefully argued by Fred Dretske (2003), we *know* we aren't zombies (and thus, we know that we are aware of things; that we see/hear/believe/want/etc things), but it is a mystery *how* we know this. As he concludes:

> We are left, then, with our original question: How do you know that you are not a zombie? Not everyone who is conscious knows they are. Not everyone who is not a zombie, knows they are not. Infants don't. Animals don't. You do. Where did you learn this? To insist that we know it despite there being no identifiable *way* to know it is not very helpful. We can't do epistemology by stamping our feet. Skeptical suspicions are, I think, rightly aroused by this result. Maybe our conviction that we know, in a direct and authoritative way, that we are conscious is simply a confusion of what we are aware of with our awareness of it. (Dretske 2003, 15)[33]

The upshot of this is that it is a mystery what information we employ to know that we are aware. If we see nothing but the blue, to use Moore's phrase, then what information do we rely on to self-ascribe that sensation as *seeing blue*? Following this line of thought, TTP is not only a problem for a *procedure* whereby we acquire self-knowledge; it is a problem about awareness itself, for our awareness doesn't seem to give us information about our mental lives.[34]

---

[31] For other descriptions of the same phenomenon, see, for instance, Harman (1990, 667) or Tye (1995, 30). For careful discussion of the phenomenon and the metaphysical conclusions drawn from it, see Kind (2003), Nida-Rümelin (2007).

[32] Following Nida-Rümelin (2007), We should be careful to distinguish the phenomenon of transparency with the assumption that the only way in which we *could* become aware of our own awareness is by becoming aware of some *new* feature of our experience. As Nida-Rümelin pointedly describes the phenomenon, when I focus on my experience of the blue 'I do not direct my attention into some inner space. I do not get aware – by attending to my own experience – of the instantiation of any property I was not already aware of before I focused attention upon my own experience' (Nida-Rümelin 2007, 429). However, this is not yet to explain which information in our experience we employ to know we *experience* it (see main text).

[33] See also Byrne (2015).

[34] This doesn't solve TTP, nor directly save transparency accounts of self-knowledge. It might even be used to maintain that self-knowledge or self-awareness is to be explained on a perceptual and reliabilist

And this, my suggestion will be, points towards a solution. If TTP connects to a problem about awareness in general, then it seems only natural to seek a solution to TTP in terms of the nature of awareness. Contrary to what is often claimed in the debate on TTP, namely that one can solve TTP by categorizing belief and judgment as distinct mental items, whether this concerns the nature of awareness of belief or of judgment doesn't seem to matter at this point. The way forward is thus, in my view, to develop an account of awareness that incorporates an attitudinal form of awareness, but that should also satisfy certain transparency requirements. For instance, that any uncertainty in one's self-ascription should direct one's attention back to the content of one's self-ascription (i.e., gazing back at the world).

model after all. However, for those who are convinced by the existing objections to such perceptual and reliabilist models (cf. Moran 2001; Shoemaker 1996), it does have positive consequences for transparency accounts of self-knowledge. The reason for this is that if TTP is a problem of conscious mentality in general, then the fact that transparency accounts face TTP is not in itself a bad thing. It might even speak in favor of transparency accounts that they bring out the problem in such an explicit form.

CHAPTER THREE

Reasoning with and without Change in Attitudes

Abstract

This paper argues against a now common analysis of reasoning in terms of mental attitudes and the (causal and rational) relations between them. According to such *attitude views*, as I will call them, reasoning is (1) a mental process that involves (2) a change in attitudes. Although reasoning often involves such a change in attitudes, e.g., forming, revising or withdrawing a belief, that doesn't imply that a change in attitudes is necessarily involved in reasoning. By discussing examples of reasoning without a change in view, it will become clear that a different approach to reasoning is needed: namely, one that includes instances of reasoning with and without change in attitudes. By combining insights from Anscombe and Frege, I will propose an alternative view of reasoning, which holds that when a person reasons she (1) makes use of conditionals, manifested in (2) a judgment of the form *p as following from q*. The paper ends by discussing the corollaries of this proposal for the relation between reasoning and mental processes in general, and a change in view in particular.

'[I]nference is something separable from the attitude of the one who is making it.'
–Anscombe 1989, 397

'Inferring is a movement of thought between propositions which may, in special circumstances, result in the thinker coming to judge the proposition inferred to be true. But no particular attitude to that proposition is implicit in inference itself, in particular not judgment of its truth.'
–Wright 2014, 28

1. Introduction

This paper argues against a now common kind of analysis of reasoning in terms of mental attitudes and the (causal and rational) relations between them. In a rather generalized way, such *attitude views* of reasoning, as I will call them, hold that reasoning is (1) a mental process that involves (2) a change in attitudes.[1] A person might first believe that there is snow outside, but then, after recognizing that it's raining, come to believe that the snow will be gone already. In the attitude view, this episode of thought is reasoning because, first and foremost, it is a mental process going from one belief to another (obviously, there will need to be additional conditions). Although reasoning often involves such a change in attitudes, e.g., forming, revising or withdrawing a belief, that doesn't imply that a change in attitudes is necessarily involved in reasoning.[2] By formulating counterexamples to the attitude view, this paper rejects that reasoning necessarily involves a change in attitudes and explores a different approach to reasoning, which includes reasoning with *and* reasoning without change in attitudes. The result of the alternative approach is also a rejection of the first element of the attitude view, namely that reasoning is necessarily a mental process. This paper thus rejects both (1) and (2) of the attitude view, although the focus of the paper is (2).

Arguing against the general idea that reasoning is a mental process that involves a change in view doesn't imply arguing against *any* involvement of mental processes and attitudes in reasoning. Rather, the alternative view I will defend, let me call it the *form view*, claims that reasoning shouldn't be *characterized* in psychological terms. It isn't the psychology that makes reasoning a recognizable phenomenon, but it is the *form* of judgment inherent in all instances of reasoning. One way to portray the difference between the attitude view and the form view is in terms of reductionism. Where the attitude view seeks to analyze reasoning in terms of smaller parts (i.e., attitudes) and the (causal and rational) relations between them, the form view explicitly rejects the possibility of analyzing reasoning in terms of smaller parts or in terms of an essential feature or property.

---

[1] See also Valaris (2018). Valaris claims that this view actually includes two distinct mental categories, which he calls deduction (a mental process) and reasoning (a change in view). I will come back to Valaris' account in section 6. Note that the second part, i.e., a change in attitudes, is often related 'to the sort of "reasoned change in view" that Harman (1986) discusses' (Boghossian 2014, 2). I have chosen "change in attitudes" because it is less committal than "change in view". Change in attitudes will be explicated in section 2.

[2] Nor that it is sufficient. Much of the contemporary debate on reasoning is focused on the question what would make a change in attitudes an instance of reasoning. The difficulty here is formulating explanatory non-circular additional conditions. I come back to this in section 4.

The form view follows a tradition that Thompson (2008) dubbed *analytic Aristotelianism*. This tradition holds that there are instances where philosophical understanding of X doesn't come from analyzing X in other terms or smaller parts, or by appealing to an essential property or feature exhibited by X (i.e., analyzing *what* X is), but from analyzing its form, i.e., analyzing the *way in which* X is.[3] This *way in which* with respect to reasoning is, according to the form view, as follows: in reasoning one makes use of conditionals, manifested in a judgment of the form *p as following from q*.

In what follows, I will first try to make more precise what the attitude view of reasoning is and especially what a change in attitudes entails (section 2). Next, I will develop a crucial argument against the attitude view. By giving numerous counterexamples, I will argue that a change in attitudes isn't necessary for reasoning (section 3). In section 4, I briefly discuss a distinct and well-recognized problem for attitude views: that formulating sufficient conditions for the attitude view invokes problems of circularity and regress. I also identify a link to reductionism. In section 5, I will explicate the form view of reasoning. I will then clarify how this alternative approach relates to reasoning with and reasoning without a change in attitudes (section 6). In this section, I also compare my view with a recent, and in various respects similar, proposal by Valaris (2018). Valaris also argues that there is an element in reasoning, which he calls deduction, that doesn't involve a change in view. In response, I outline that what Valaris calls reasoning and what he calls deduction both presuppose the form of judgement identified in the previous section. Section 7 states the conclusion.

2. Reasoning as change in attitudes

The current orthodoxy in the philosophy of reasoning is to regard reasoning as a psychological process and analyze it in terms of the mental attitudes involved and the relations between them. For instance, Broome (2013, 221) writes that 'reasoning is a process whereby some of your attitudes cause you to acquire a new attitude.' McHugh and Way (2018, 167), too, state that in reasoning '[y]ou bring some existing attitudes to mind, saying their contents to yourself, and make a kind of transition to a further attitude which you thereby acquire.' And Boghossian (2014, 2) writes that '[b]y "inference" I mean the sort of "reasoned change in view"

[3] Cf. Thompson (2008, 11); Hlobil & Nieswandt (2016, 182). See also, Boyle (2005); Ford (2015); Frey (2013); Vogler (2001). See also Valaris (2018) for a similar approach to reasoning.

that Harman (1986) discusses, in which you start off with some beliefs and then, after a process of reasoning, end up either adding some new beliefs, or giving up some old beliefs, or both.'[4] In a rather generalized way, these attitude views thus hold that reasoning is (1) a mental process that involves (2) a change in attitudes. All adherents of this view of reasoning admit that (1) and (2) aren't sufficient conditions of reasoning, but they do hold that they are necessary conditions.

It seems quite plausible that reasoning often involves a change in attitudes. If you care to know whether there are any beers left (because perhaps you want one) or whether the snow is melting (because you want to make a snowman) or whether the streets are wet (because you want to go roller-skating), you might reason as follows:

(1) If Jane had a beer, then there are none left. Jane had a beer. So, there are none left. (McHugh & Way 2018, 167)
(2) If it rains, the snow melts. It is raining. So, the snow melts. (Broome 2013, 216)
(3) If it rained last night, the streets are wet. It rained last night. So, the streets are wet. (Boghossian 2014, 2)

In these cases, you adopt a belief in the conclusion, e.g., that there aren't any beers left, that the snow melts, and that the streets are wet, and thus you change your attitudes.

To go beyond such initial plausibility, we first need to understand what a change in attitudes is. I provide three clarifications. First, the attitudes in question won't include merely entertaining that *p*, thinking about *p* or supposing that *p*. I will assume that these ways to think about *p* don't constitute an attitude vis-à-vis *p*.[5] Rather, I follow a common way to portray mental attitudes: having a mental attitude regarding *p* involves specific commitments or a specific stance towards *p*, where *p* is represented as being a certain way, i.e., true, false, valuable, etcetera (cf. Burge 1998). The account I will thus assess is:

[4] "Inference" and "reasoning" are used interchangeably in this paper.
[5] If one would include these latter kinds of thinking about *p* in one's category of mental attitude, then it would lead to an attitude view that might be true, but only trivially so: that reasoning involves thinking (whether we call it entertaining some proposition, supposing it, or just thinking) about the topic under consideration is so uncontroversial that it becomes an insubstantial claim and one not meriting much evaluation (cf. Valaris 2018). It might be worth noting that most attitude accounts start with a quite strong notion of attitude but sometimes stretch the notion of attitude when faced with counterexamples. This is a response that will play a role in the discussion of example (iv) in section 3.

> Attitude view: Reasoning is (1) a mental process involving (2) a change in stance towards the object of one's thought.

A second question regarding the nature of a change in attitudes is which attitude is supposed to change. Which attitude related to the inference is the one that constitutes a change in attitudes? If we take another look at the quotations in section 2, we see that the change in attitudes takes place in the attitude vis-à-vis the conclusion of one's reasoning: changing your attitudes is to 'acquire a new attitude' (Broome), 'transition to a further attitude which you thereby acquire' (McHugh & Way), or 'adding some new beliefs, or giving up some old beliefs' (Boghossian). Moreover, the examples (1)-(3) also indicate that the change in attitudes concerns the conclusion, for, in each of these examples, you adopt a belief in the conclusion. Hence, acquiring trivial beliefs such as the belief that you just went through an inference, doesn't count as a change in attitudes relevant for the process of reasoning. A change in attitudes, then, consists in a change in stance *regarding the conclusion* of one's reasoning. Hence, the change in attitudes in the attitude view should be: (2) a change in stance regarding the object of the conclusion-thought.

A last question to be addressed is what kind of change in attitudes suffices for a change in attitudes. For simplicities sake, let's focus on belief (which I will do for the remainder of the paper). Again, there are two options, following the distinction between full belief and degrees of belief (credences). A change in attitudes may require either a change in credences (e.g., from 0.7 credence to 0.9 credence) or a change in full belief (e.g., from belief to disbelief or to suspension of judgment). The difference between the two is that only in the latter case a change in attitudes consists in the adoption, revision, or withdrawal of a belief. Given these options, there are two possible ways to explicate a change in attitudes in the attitude view.

> The credence attitude view: Reasoning is (1) a mental process involving (2) a change in credences regarding the object of the conclusion-thought.

> The full attitude view: Reasoning is (1) a mental process involving (2) a change in full belief regarding the object of the conclusion-thought.

Both accounts will be under discussion in the next section.

3. Counterexamples to the attitude view

Does all reasoning involve a change in attitudes? There are several examples that seem to suggest otherwise. In this section, I will discuss six counterexamples: (i) sustaining belief in the conclusion, (ii) Knorpp's example of puzzle-solving, (iii) non-formation of belief, (iv) hypothetical reasoning, (v) reasoning with an incoherent premise, and (vi) interpersonal reasoning. The only example where proponents of the attitude view might have a response is hypothetical reasoning, if the attitude view includes not only attitudes regarding the actual world, but also regarding possible worlds. Still, the other examples count against both attitude views.

(i) The first example is about sustaining belief in a conclusion. Since sustaining a belief in the conclusion might involve a change in credence of belief, one might suppose that it brings out a difference between the credence attitude view and the full attitude view. But, in fact, it is a counterexample to both accounts. The full attitude view may be held untenable for a very simple reason: we often reason towards a conclusion we already believe, for instance, when we recognize other reasons pertaining to the same conclusion. Imagine, for instance, that after reasoning through example (2) and adopting the belief that the snow will melt, I reason through the following example:

> (4) If the temperature rises above zero degrees Celsius, then the snow will melt. The temperature rises above zero degrees Celsius. So, the snow will melt.

Since you already believe that the snow will melt, you cannot, in reasoning through (4), adopt a *new* belief that the snow will melt. Hence, this is a strong case against the idea that a change in attitudes necessarily involves a change in full belief. And so it seems that a change in attitudes should involve, minimally, a change in credence of belief. It might be the case that, after reasoning through (4), one's credence in the belief that the snow will melt increases.

However, a similar case against the credence attitude view is also easily conceivable, i.e., one where one's credence in belief doesn't change after going through another piece of reasoning leading to the same conclusion. What if one goes through another piece of reasoning just for pragmatic reasons, e.g., for the fun of it? Or for aesthetic reasons, e.g., that one feels there must be an easier way to arrive at

the same conclusion? Consider a mathematician who has just formulated a proof for a specific theorem and knows that the proof is the same as the one formulated by many mathematicians before her who tried to find a mathematical proof for this theorem. Still, she cannot shake off the feeling that there must be a simpler proof possible. Note that this feeling concerns the way in which the conclusion is arrived at and not the conclusion itself. She has full credence in the truth of the theorem. Subsequently, she sets out to formulate a simpler proof for the same theorem. After a significant amount of time, she has found a new way to arrive at the same theorem. In all likelihood, it seems that she has reasoned, but without changing her belief or the credence of her belief in the conclusion. Hence, she has reasoned without changing her attitudes. This is a case against both the credence attitude view and the full attitude view.

(ii) The second example concerns reasoning where one doesn't arrive at a conclusion yet. Knorpp (1997, 81-2) provides a suitable example, namely working out *The Riddle of Dracula*. Put in brief terms, the puzzle is the following (adapted from Knorpp): there are four groups of people in Transylvania, determined by two variables. Each person is either a human or a vampire and either sane or insane. Humans always tell the truth, vampires always lie. Sane people believe all and only true propositions, insane people believe all and only false propositions. Question: what one question can you ask of a Transylvanian which, given that he answers 'yes' or 'no', will allow you to ascertain whether Dracula is alive or not? As Knorpp describes (and as you are likely to experience too, if you were to try to answer this question), on his first try he thought (i.e., reasoned) about this puzzle for twenty minutes, failing 'to arrive at a solution of any kind' (ibid.). On Knorpp's second try, he did arrive at a conclusion (and so, we might say, his attitudes changed). Whereas on Knorpp's second try he did change his attitudes, on his first he didn't. Still, 'it would be terribly implausible to call what [he] was up to anything other than "reasoning".' Reasoning without arriving at a conclusion is a case against both attitude views.

(iii) Imagine a case where a person, say Elisabeth, reasons through (1), i.e., from the premises that Jane had a beer, and that if Jane had a beer, then there are none left, to the conclusion that there are no beers left. However, Elisabeth is really tired at the moment, so she actually doesn't adopt the belief that there are no beers left. We know this, because soon afterwards she gets up to grab a beer only to find out (again) that there are none left. According to both attitude views, Elisabeth's failure to actually change her attitudes is ground to deny that she has reasoned. But this

seems absurd. A rational failure to adopt a belief (or change one's credence in belief) in the conclusion shouldn't be ground to decide whether the episode of thought one just went through was a piece of reasoning or not. One can reason even if one fails, due to certain irrational influences, to adopt a belief or change one's credence regarding the conclusion.

(iv) The fourth example concerns hypothetical reasoning. One can reason through examples (1)-(3) from the introduction in order to know whether there is any beer left, the snow is melting, or the streets are wet, but one can also reason through them in a hypothetical manner. Suppose that if Jane had a beer, then there would be none left, and that Jane had a beer. Then, there would be none left. Or suppose that it rained last night and that if it rained last night, then the streets would be wet. Then the streets would be wet. What should we say about these examples? Do they involve a change in attitudes? This is a difficult issue and takes some time to spell out.

Intuitively, hypothetical reasoning is precisely the kind of reasoning that brackets the question of whether one believes the premises and conclusion. The expression "for the sake of argument" is precisely to do just that: to bracket one's mental attitudes to the topic under consideration, i.e., one's commitments to the truth or falsity of the propositions involved in the inference. This implies that one's mental attitudes towards the propositions involved are *irrelevant* in the case of hypothetical reasoning. Consider the following example from Valaris (2018, 4):

> I might…consider the hypotheses that God is omnipotent, omniscient, and infinitely good and that evil nevertheless still exists, and see what follows. I may do this while lacking or suspending any attitudes towards the original hypotheses, and without any disposition to adopt any particular attitude towards any consequences I deduce from them. … [W]hat attitudes the agent has towards her hypotheses is *irrelevant* to [hypothetical reasoning]: a theist, an atheist and an agnostic may deduce exactly the same consequences, and in the same way, from the original hypotheses, even if they take incompatible attitudes towards them.[6]

[6] Note that Valaris is here explaining his notion of deduction, which he distinguishes from reasoning. I will come back to the notions that Valaris uses (or, as he is careful to note, stipulates) and the distinctions he makes in section 6.

Thus, hypothetical reasoning, or reasoning where one is interested in some consequences of a set of claims, 'does not appear necessarily to involve – much less to *consist in* – adopting or revising any such attitudes' (ibid.).

There are two relevant ways of responding to this challenge to the attitude view. One option is to deny the relevance of hypothetical reasoning altogether. According to this response, hypothetical reasoning indeed doesn't involve a change in attitudes, but this is how it should be. Hypothetical reasoning is, according to this response, just the same as, say, an inference written down on paper, or an argument in a textbook. This response claims that since hypothetical reasoning only concerns relations between contents, it isn't a form of reasoning proper. This is one of the lines of response that McHugh and Way (2018) propose. They side here with Harman (1986), who distinguishes between the category of argument (or logic) and the category of reasoning as a psychological process. The problem with this response is, in my view, that it begs the question. First, this response assumes that reasoning is a mental process and thereby adopts the distinction between reasoning as a psychological process and the logical aspects of inference. Next, when faced with the difficulty that hypothetical reasoning is a form of reasoning that cannot be characterized as a psychological process involving a change in attitudes, Harman's distinction is simply repeated to argue that hypothetical reasoning thus belongs in the category of logic. However, without presupposing Harman's distinction, it seems quite plausible to regard hypothetical reasoning as a form of reasoning without a change in attitudes. After all, hypothetical reasoning is something we *do*; it manifests itself in a person thinking certain thoughts. This makes it different from sheer logic or proof, which are items that also exist without a person doing anything. The first response thus begs the question, because it presupposes that all reasoning that doesn't involve a change in attitudes belongs in the category of logic.[7] This is only plausible by having already assumed their own definition of reasoning in the first place.

The second option is to say that hypothetical reasoning does involve a change in attitudes, it is just different from the normal case. One might accept that, indeed, hypothetical reasoning doesn't involve any mental attitudes that purport commitments to the actual world being a certain way (let's call them categorical attitudes), but it might still involve attitudes that purport commitments to *possible worlds* (let's call them hypothetical attitudes). Accordingly, the change in attitudes in hypothetical reasoning concerns the adoption of a belief in a conditional

[7] See Knorpp (1997, 82-8) for further exposition of Harman's argument and for critical analysis of Harman's argument that logic isn't relevant to belief-revision.

conclusion.[8] Let's try to make more precise what this entails by looking at example (2) about the rain and the snow. Whether one reasons categorically or hypothetically, the inference itself is the same, namely from:

(p1) If it rains, the snow melts.
(p2) It is raining.

…to:

(c) The snow melts.

That is to say, in both instances of reasoning, there isn't a logical difference in the inference.[9] Still, one might say that the difference is that in the hypothetical case the conclusion drawn isn't *c* itself but a conditional statement: *if p1 and p2, then c*. This would mean that one adds a new conditional, or new truth-connection, to one's belief set.

I think this response has certain merits. Hypothetical reasoning is often used to find out about conditionals and which of them should or shouldn't be believed. Still, I don't see why we should accept that hypothetical reasoning necessarily involves adopting a belief in a conditional conclusion. First of all, examples (i)-(iii) also count against this view of hypothetical reasoning. Sustaining belief in a conditional conclusion, reasoning without coming to a conclusion, and failing to adopt a belief in a conditional conclusion are counterexamples to the idea that hypothetical reasoning necessarily involves a change in attitude regarding a conditional conclusion too. But secondly, I will add two counterexamples that specifically address instances of hypothetical reasoning.

(v) One can suppose, "for the sake of argument," premises that are incoherent. One can work out what follows from incoherent premises, but one will not adopt any belief in a conditional conclusion. Hence, this is a counterexample to the idea that hypothetical reasoning necessarily involves the adoption of a belief in a conditional conclusion.

[8] Cf. Broome 2013; McHugh and Way 2018.
[9] One might think that the logical identity between categorical and hypothetical reasoning is reason to include mental attitudes in one's account of reasoning. That is the only way in which they can be distinguished. However, that including mental attitudes is the only way to distinguish between the two doesn't imply that understanding the attitudes involved provides us with an understanding of the nature of reasoning. Nor with an understanding of why both cases are an instance of this same phenomenon, namely reasoning.

(vi) The last example, which I call *interpersonal reasoning*, is an instance of (iv) and based on an example by Anscombe (1989, 395), which I have adapted to suit theoretical reasoning. The example is about two persons, where person A provides the premises and person B draws a conclusion from the grounds set forth by person A:

> Suppose I say to you: [premise 1:] "You live in a democracy." [Premise 2:] "If you live in a democracy, you should take responsibility." And suppose I then give you a prudish look with nothing more said, whereupon you think "Yeah, yeah, so I should take responsibility."

The idea of this example is that you (person B) draw the conclusion from the two premises I (person A) set forth. In order to make the inference, person B must think about all the elements of the inference (the propositions), but it doesn't require her to present the propositions in a certain way. That is to say, she doesn't need to have any attitudes towards the premises or conclusion to be able to "reason along" with person A.

This also holds, interestingly enough, for the conditional statement "If you live in a democracy, you should take responsibility". Making an inference does not require person B to believe the conditional, nor need it be true. How can that be? Suppose person B knows that there are multiple exceptions to the conditional statement. For instance, children who live in a democracy shouldn't take responsibility, they should be taken care of. Hence, the conditional is false and person B knows that it is false. Still, the reasoning is *comprehensible* to her. Intuitively, she *can* still draw the conclusion that she should take responsibility. And she can do this despite the fact that she disbelieves the conditional statement postulated by person A.[10]

Hence, again we have an example of a piece of reasoning without a change in attitudes. So, reasoning doesn't necessarily involve a change in attitudes. And neither does hypothetical reasoning.

To conclude, examples (i)-(iii), (v) and (vi) are counterexamples to both attitude views, because they are examples of reasoning without change in (credence of) belief. Only example (iv), hypothetical reasoning, isn't a clear-cut counterexample to both attitude views. What is clear is that in hypothetical reasoning one's

[10] What's more, it seems she can also draw the conclusion while at the same time disbelieving it. But again, her belief or disbelief in the conclusion isn't something that *follows* from going through the inference.

categorical attitudes are irrelevant. That is, whether one believes, disbelieves or suspends judgment on the premises and conclusion doesn't matter for the course of one's reasoning nor for the fact of whether one is engaged in reasoning (or, perhaps, some other kind of thinking). Still, one might understand hypothetical reasoning as involving hypothetical attitudes instead of categorical ones. But even if hypothetical reasoning might often lead to the adoption of a belief in a conditional conclusion, there is no reason to accept that this is a necessary condition of hypothetical reasoning, as shown, for instance, in example (v) and (vi), but also in (i)-(iii). Hence, a change in attitudes isn't a necessary condition of reasoning.

The alternative is to view a change in attitude not as a necessary condition of reasoning, but as a specific kind of reasoning. In this view, reasoning in itself can be understood without reference to the reasoner's attitudes towards the premises and conclusion. This alternative view seems to be reflected in Anscombe's statement that "inference is something separable from the attitude of the one who is making it" (1989, 397). And, relatedly, in Wright's claim that we should "distinguish inference in general from *coming to a conclusion*…; no particular attitude to [a] proposition is implicit in inference itself" (2014, 28). Anscombe and Wright, in other words, claim that reasoning that involves a change in attitudes is but one instance of the more general phenomenon that reasoning, or making an inference, is.

Importantly, denying that reasoning always involves a change in attitudes leaves much common ground in different views of reasoning unaltered. Proponents and adversaries of the attitude view consider reasoning as something we *do*; as a person-level activity in thought; and as something that is a conscious activity (which, as we will see later, need not necessarily be a mental *process*). Moreover, both sides agree that when a person reasons, she thinks certain thoughts and thus that, in a sense, reasoning depends on her psychological constitution. The point of disagreement is whether the involvement of a person's psychological constitution implies that reasoning should be characterized in psychological terms: that is, whether the psychological constitution plays a constitutive or an enabling role.

4. Other considerations against the attitude view

The previous section argued that, given numerous counterexamples, a change in attitudes isn't a necessary condition of reasoning. As also mentioned previously, neither is it a sufficient condition of reasoning. This is, in and of itself, unproblematic. However, in trying to formulate additional conditions, attitude views run into problems of circularity and regress. In this section, I will illustrate

this by reference to Boghossian's Taking Condition. Moreover, I will draw attention to the reductionist tendencies inherent in attitude views. Neither of these things is central to the argument developed in this paper, but they help motivating and situating the account that I will propose in the next section.

It is clear that merely having a change in attitudes isn't sufficient for reasoning. For instance, if one forgets that there is a new mayor in one's hometown, one's attitudes are changed, but one didn't reason. This means that an account of reasoning that starts from the idea that reasoning involves a change in attitudes needs to formulate additional conditions that a change in attitudes must satisfy for it to be reasoning. Basically, all of the current debate spirals around formulating such additional (and sufficient) conditions. Moreover, it is widely recognized that it is difficult to formulate such conditions in a non-circular manner or without facing problems of regress.[11]

As an illustration of this point, consider the Taking Condition as formulated by Boghossian (2014, 5):

> Taking Condition: Inferring necessarily involves the thinker *taking* his premises to support his conclusion and drawing his conclusion *because* of that fact.

The Taking Condition is supposed to ensure that the causal relation between one's beliefs in the premises and one's belief in the conclusion isn't merely causal but is of the right (non-wayward) and thus rationalizing kind. Boghossian's Taking Condition seeks to secure this by introducing another mental item, i.e., "the taking", on top of the thinker's attitudes regarding the premises and conclusion. Given that such a "taking" is an additional item, the question arises which role this item is supposed to play in reasoning. In trying to account for the role of the taking, one inadvertently seems to run into regress problems. First, the role of taking in the inference shouldn't be that of an additional premise, as is familiar from Carroll's argument (1895). Moreover, its role shouldn't be merely causal, but it should rationalize the inference. However, if it is to rationalize the inference, it seems unavoidable that the content of the taking should be related, by the thinker herself, to the content of the inference, and as of yet there appears to be no way to relate the

[11] Cf. Boghossian 2014; Broome 2013, 2014; Hlobil 2014; McHugh & Way 2016, 2018; Valaris 2016, 2017; Wright 2014; among others. The debate focusses predominantly on the impossibility of giving a non-circular account of rule-following, of the causal relation involved in reasoning (and how it is of the right non-wayward kind), and of the taking condition. Space forbids me to go into the circularity and regress problems in detail, but they are well-documented in the aforementioned literature.

two without any form of inference. Hence, the Taking Condition condemns accounts of inference to problems of regress.[12]

The claims that a change in attitudes isn't sufficient for reasoning and that formulating additional conditions runs into problems of circularity or regress are both widely recognized in the debate, but this recognition doesn't seem to lead to a reevaluation of the idea that reasoning should be characterized as involving a change in attitudes. One motivation for this steadfastness might be that this approach fits in the currently dominant scientific view of the mind. This may function as a kind of external justification for this approach (cf. Rödl 2007, 209; Velleman 2000, 129). Analyzing reasoning in terms of mental attitudes and relations between them gives us a philosophical picture of reasoning that can be used in the psychology and neuroscience of reasoning too. The attitude view fits in the current dominant, reductionist approach to the mental realm. It seems almost peremptory to analyze mental phenomena (or phenomena involving the mind), such as self-knowledge, intentional action, and reasoning, in terms of mental attitudes (and the relations between them). For instance, it is often presupposed that self-knowledge of a mental attitude concerns a second-order belief about a first-order attitude. As such, self-knowledge becomes a matter of designing the right relation between second-order belief and first-order attitude. Similarly, it is often assumed that understanding intentional action means understanding the right relation between an intention and behavior. And as we have seen in this paper, on the leading philosophical view of reasoning, reasoning is analyzed in terms of moving from premise-beliefs to a conclusion-belief. The result might be called a reductionist approach, not because the mental phenomena are reduced to one component or feature entailed in the phenomenon or to a different explanatory level (e.g., psychological, neurological), but because the approach assumes that a correct philosophical analysis of these phenomena starts by analyzing the phenomenon under discussion in terms of smaller parts and the relations between them.

A reductionist analysis is, of course, not bad in itself, nor is being motivated by providing a philosophical analysis of reasoning that makes it a topic suitable of scientific investigation. However, in the case of reasoning it seems to lead to a view of reasoning that, first, doesn't correspond to all instances of reasoning and thus not to the phenomenon that reasoning is, and secondly, faces problems of circularity and regress. Should we accept these problems? Not if there is another approach

[12] This is an extremely short review of the route from the Taking Condition to regress. For more in depth analysis, see, for instance, Boghossian (2014); Broome (2013, Ch. 12); McHugh & Way (2016); Valaris (2014).

available. The problems, at least, provide grounds to reconsider the orthodox (and reductionistic) view of reasoning.

5. The form view

What kind of approach to reasoning would include both instances of reasoning, i.e., with and without a change in attitudes? The alternative approach of reasoning that I want to sketch in this section combines insights from Anscombe and from Frege. This so-called form view holds that when a person reasons, she (1) makes use of conditionals, manifested in (2) a judgment of the form *p as following from q*. Before explicating both elements in turn, a qualification is in order. The ideas presented in this section aren't yet full-fledged, but I hope to convince the reader that the approach itself looks promising.

Some recent accounts of reasoning seek to define reasoning as an activity with one specific aim. For example, McHugh and Way (2018, 178) argue that 'the ultimate point of reasoning is to get fitting attitudes. In other words, it is to *get things right*.' Valaris (2017, 2016), by contrast, claims that the aim of reasoning shouldn't be characterized on the attitudinal, and what he calls *syntactic* level, but on a *semantic* level: 'the epistemic aim of reasoning is to reduce uncertainty about the world, *via* the elimination of alternative ways the world might be.' But if we look at all the different ways and different contexts in which we reason, and specifically, if we look at examples (i) – (vi), then it seems that reasoning has many different aims. We don't merely reason *to get things right* or to reduce uncertainty about the world, but also for the fun of it, to explore new possibilities, to open new possibilities (cf. Kompridis 2000, 293), to determine what to do, investigate, etcetera.

Where others seek to identify one specific aim of reasoning as its determining feature, I think we should do justice to the manifold goals with which we reason. To take up a suggestion made by Anscombe (1989) in her chapter on "Von Wright on Practical Inference," we might say that reasoning is a way of using a specific kind of tool. Reasoning is to put "implications" or "truth-connections" between propositions to a particular service.[13] This means that, even if reasoning doesn't have one specific aim, it does have a *point*: to drag out implications.

Anscombe clarifies her argument with the following example about plant growth (1989, 394):

[13] What it is to put X to a particular service requires more detailed analysis and might require a different analysis in the case of theoretical and practical reasoning. However, this doesn't impinge the general point about reasoning. Cf. Müller (1979).

(1) If these substances are in the soil, the plants will be fed by them. (if r then q)

(2) If plants are fed with certain substances, there will be spectacular plant growth. (if q then p)

These implications might be put to different use. For instance, if it is given or assumed that these substances are in the soil (r), one should, in accord with (1) (if r then q) and (2) (if q then p), come to the (assumed) conclusion that there will be spectacular plant growth (p) (theoretical reasoning). Or, if one is to investigate why there is spectacular plant growth (p), one should, according to (2) (if q then p) and (1) (if r then q), examine the soil to check whether those substances are present (investigation). Again, these same considerations might figure in practical reasoning: if the objective is to attain spectacular plant growth (p), then, given (2) (if q then p) and (1) (if r then q), one should (decide to) put those substances in the soil (r). Anscombe's formalizations might be of help here (1989, 393):

Theoretical reasoning	**Investigation**	**Practical reasoning**
r	Given: *p*	Wanted: that *p*
if *r* then *q*	if *q* then *p*	if *q* then *p*
if *q* then *p*	if *r* then *q*	if *r* then *q*
p	To investigate: *r*	Decision: *r*!

In each instance of reasoning, the 'considerations and their logical relations are just the same' (Anscombe 1989, 392). In whatever way one reasons, one makes use of conditionals.

But what does "making use of conditionals" entail? This question brings us to the second element of the form view: making use of conditionals is manifested in (2) a judgment of the form *p as following from q*. Making use of conditionals is to make a particular kind of judgment. In order to explain this form of judgment and why it is a genuine alternative to the attitude view, I want to return to Boghossian's Taking Condition. Boghossian's condition is inspired by the following statement of Frege:

> To make a judgment because we are cognizant of other truths as providing a justification for it is known as *inferring*. (Frege 1979, 3)

Boghossian interprets this as saying that '[a] transition from some beliefs to a conclusion counts as inference only if the thinker *takes* his conclusion to be *supported* by the presumed truth of those other beliefs' (2014, 4). This interpretation leads him to his formulation of the Taking Condition, which postulates "the taking" as an additional mental item involved in reasoning, with regress problems as a result.

Boghossian's interpretation of Frege's statement, however, isn't uncontentious. Frege doesn't mention any "taking" nor moving from premise-beliefs to a conclusion-belief. What Frege does mention is what must be true of a particular judgment in order for it to be a case of inferring. In contrast to Boghossian's interpretation, what Frege is doing here could well be interpreted as describing what reasoning *is*, namely a specific kind of judgment. If one's judgment is such that one makes an inference, then one makes a judgment in virtue of it being supported by other (presumed) truths. If one infers that it is raining from seeing drops in the puddles outside, then one judges that it is raining *as following from the truth of* there being drops in the puddles outside. Put briefly, if one infers *p* from *q*, then one judges that *p as following from q*. On this interpretation, Frege doesn't describe a process or an additional mental state. Rather, he states what kind of judgment is involved in reasoning.

One might be inclined to think that claiming that reasoning consists of a specific form of judgment comes down to claiming that this form of judgment is a necessary condition of reasoning. But if it would merely be another proposal of a necessary condition of reasoning, the approach would run into the same problems of circularity and regress as the attitude view. The reason for this is that it doesn't give us a non-circular understanding of the nature of reasoning. We can see this by asking the following question. What kind of *following* is at issue in a judgment of the form *p as following from q*? Surely, it is the kind of following where the truth of *q* supports the truth of *p*, and not just a causal or temporal sequence between *q* and *p*. But saying that the kind of following we are after is a conditional is just to say that the person is reasoning and not memorizing a temporal sequence. In short, to understand a judgment of the form *p as following from q* is just the same as understanding what reasoning is. Hence, the form of judgment explicated in Frege's statement shouldn't be understood as a necessary condition of reasoning. But can it mean anything else?

An alternative interpretation is available, but it takes some time to spell out. Interpreting Frege as describing a form of judgment is based on a philosophical approach that has been dubbed *analytic Aristotelianism* (cf. Thompson 2008). The best-known exemplar of this tradition is Anscombe's monograph *Intention*. One key

claim in this book is that 'the term "intentional" has reference to a *form* of description of events' (1957, §47). Importantly, the reference to form is not meant to depict an essential *feature* or *property* of intentional actions. Anscombe explicitly denies that understanding the nature of intentional action can be a matter of analyzing it in terms of a specific feature or property, or by stating necessary and sufficient conditions.[14] She denies this because she thinks that any such analysis will run into circularity problems: it will need to presuppose some understanding of what we mean when we call an action intentional.[15] This is most obvious in her remarks on the why-question that asks for a reason for action. Anscombe claims that an intentional action is an action subject to such a why-question. This has led many to conclude that Anscombe's view is that an intentional action *is an action done for a reason*. However, Anscombe is careful to note not only that occasionally intentional actions are done for no reason at all, but also that *done for a reason* is not giving us any more information about what an intentional action is. For, if we want to understand what *kind* of reason we mean, and distinguish it from a mere causal reason, then we need to presuppose the same distinction that we are trying to understand. '[W]e should be going round in circles,' as Anscombe (1957, §5) writes. Hence, the reference to "form" discloses a completely different approach to intentional action: not describing a property of intentional action but, what is now called, its *logical form*. My suggestion is that we should also understand Frege's form of judgment as grasping the logical form of reasoning. Let me therefore explicate what analytic Aristotelianism entails and what the notion of logical form means.

Analytic Aristotelianism has its roots in Aristotle and in the analytic tradition, especially in the work of Frege. Analytic about this approach is that it focuses on the *logical* form and Aristotelian about this approach is that it focuses on so-called "form concepts." Let me address both points in turn, starting with the latter. "Form concepts" are concepts that unite a class of things, say Xs, not because of specific properties that Xs bear, but because Xs have a certain *form* (cf. Thompson 2008, 11). Hence, "form" in this context isn't a property (more on this in the next paragraph on what makes this form *logical*).[16] 'Philosophical comprehension of the concepts in

[14] For recent illuminating papers on Anscombe's method, see Ford (2015); Frey (2013); Hlobil & Nieswandt (2016); Vogler (2001).

[15] This is also the most central problem faced by attitude views of reasoning, as, for instance, Boghossian (2014) himself also points out.

[16] As an illustration, consider self-knowledge of my intention. Suppose I know that I intend to go to the movie's tonight. Moreover, you know this too. Hence, we know the same fact about me, namely that I have this intention. Still, this same fact plays a completely different role for us, not least because it makes no difference to how you continue with your day, whereas for me it does: I have to actually make it happen that I go to the movies tonight. Your knowledge and my self-knowledge aren't distinctive because it involves different information, but it is different because the information plays a

question,' as Thompson (ibid.) writes, 'will come from grasping the specific character of this form of unity in each case.' This means that such philosophical understanding doesn't come from analyzing the unity in other terms or smaller parts, or by appealing to an essential property or feature exhibited by X. Rather, this approach holds that definition – in terms of providing 'informative and non-circular necessary and sufficient conditions' or in terms of an appeal to a specific property – is not the only mode of explanation (cf. Hlobil & Nieswandt 2016, 182). This approach thus holds that explaining *what* something is, is not the only route to philosophical comprehension. Sometimes, i.e., in case of "form concepts," philosophical comprehension requires explaining *the way in which* something is – viz. its *mode of being* or the *kind* of thing it is (cf. Boyle 2005).

The distinctions drawn by the concepts under discussion are, in line with the analytic tradition, not natural but *logical* in character. Importantly, "logical" in this tradition doesn't refer to logical principles or to answering to such principles. Rather, it means that we are not after any kind of form, but a *form of judgment* or a *form of thought*. Analyzing the form of the unity depicted by the concept is to lay bare its logical structure or its structure in thought. Most of the time, the logical structure of thoughts or concepts refers to their formal character, not to their content. However, the focus here is the other way around, namely on content and not on formal characteristics. The idea is that there are some quite particular concepts that are *contentful* but can still only 'be comprehended precisely through a reflection on forms of thought or judgment' (Thompson 2008, 14). In Frege, some of these concepts are "number" and "concept." In Thompson, the relevant concepts are "life-form," "intentional action" and "practical disposition." These concepts are the "form concepts" – concepts that depict *the way in which* something exists. The logical structure of such concepts is revealed by analyzing the things that can be said or asked about X, and thus by analyzing our practices and abilities regarding X.[17]

Form concepts thus unite a class of things, not in virtue of specific properties they bear, but in virtue of the form they have – i.e., the kind of things they are. Such a form refers to what can be predicated of the thing in question, which is to say that the form under consideration is a logical one. Hence, analytic Aristotelianism seeks philosophical understanding of a concept by analyzing its logical form. This is a non-reductive approach since reasoning isn't analyzed in smaller parts or necessary and sufficient conditions but as a "contentful form concept." As such, the concept

different role. My tentative suggestion is that we should also understand this as a distinct logical form of knowledge. (cf. Wittgenstein 2009 [1953]; Moran 2001; Boyle 2019). See also the Concluding Reflections.

[17] cf. Boyle (2005, 2009a); Ford (2015); Hlobil & Nieswandt (2016).

reasoning unites a class of things, not because they share particular *features*, but because they bear a similar logical structure, i.e. a similar form of judgment.

I hope this exposition of the approach and of the concept of logical form suffices to show that Frege's form of judgment need not be understood as a specific feature or necessary condition of reasoning. Such understanding, after all, would just make us "going round in circles": judging that *p as following from q* doesn't provide us with an analysis of reasoning in terms of something else, but explicates the *form* that reasoning has. Frege's form of judgment reveals a structure inherent in all the things that seem to be united under the concept reasoning. Whether a person is drawing up an argument, solving a puzzle, trying to follow someone else's line of reasoning, deliberating about what to believe, she judges that something follows from something else. And whether a person sees a truth-connection immediately or needs some time to imagine all the different possibilities before seeing it, she judges that something follows from something else. It thus seems fruitful to take reasoning to be a form concept.

To conclude, the form view parts ways on two central points in the current debate on reasoning. First, when we reason we aren't always after the truth or after reducing uncertainty. Rather, we can use reasoning for many different aims. This is ground to conceive reasoning, not as an activity with an essential aim, but as putting a tool to use, and this tool consists of conditionals. We can thus say that the *point* of reasoning is to drag out implications. Secondly, making use of conditionals is manifested in a form of judgment, namely a judgment that *p as following from q*. This results from interpreting Frege's statement from an analytic Aristotelian approach. And it diverts from Boghossian's interpretation that Frege describes a judgment (a taking) that is involved in reasoning in addition to the premise-beliefs and conclusion-beliefs. The resulting view is explicitly non-reductionist because it doesn't analyze reasoning in terms of smaller parts but seeks to understand its logical form. It doesn't give a characterization of reasoning in terms of *what* it is, but in terms of the *kind* of thing it is.

6. Reasoning with and without change in attitudes

The previous section still leaves open why the alternative approach is suitable both for reasoning with and reasoning without a change in attitudes. Moreover, it doesn't speak to implications about the relation between reasoning and mental processes. Hence, let's draw some corollaries in this section.

If reasoning is characterized as making use of truth-connections, embodied in a judgment of the form *p as following from q*, then it is simply irrelevant whether mental attitudes (categorical or hypothetical) are involved. What matters is not whether a specific conclusion is *believed*, but whether the thought or judgment involved has a specific form. This means that the involvement of mental attitudes isn't, as is the case in the attitude view, constitutive of reasoning. Rather, it is the other way around. A change in attitudes can be a consequence of many different things, such as perception, forgetting, remembering, a bump on the head, and *also of reasoning*. What makes the case of reasoning distinct from these other cases of a change in attitudes is that a judgment of the form *p as following from q* is involved. It is this judgment that makes a change in attitudes an instance of reasoning.

As mentioned before, reasoning often results in a change in attitudes. When we seek to determine what to believe, do, value, investigate, etcetera, our attitudes will change in the course of reasoning. If a person believes that *q* and then makes the judgement *p as following from q*, this normally means that she will then also believe that *p*. That is to say that a person who believes (or wants, etc.) the premises, will, when she reasons, normally also believe (and do etcetera) the conclusion. I say "normally," because there may be irrational (and perhaps also a-rational) factors that influence the adoption of a new belief.[18] But again, the adoption of the belief itself or failure thereof doesn't indicate whether the person was or wasn't reasoning.

As a consequence, we should distinguish between the logical and psychological aspects of reasoning, but in a different way than on the attitude view, where the distinction is drawn from Harman (1986). He makes a distinction, as already mentioned in example (iv), between the category of argument and the category of reasoning (identified as psychological process). In my view, the psychological and the logical aspect part differently: if one wants to discuss a piece of reasoning, question it, check it, determine whether it is good, then one engages only with the content, i.e., with the truth-connections between the propositions. By contrast, if one is interested in how certain mental attitudes and mental processes are informed (psychologically) by those connections, in the history of someone's mind, then we should include mental attitudes in our description (cf. Vogler 2001, 33-7). But being able to chronicle such a history as an *episode of reasoning* doesn't depend on which mental attitude caused another, but on how the content of those states is informed (psychologically) by truth-connections between propositions.[19]

---

[18] For instance, if one learns of something hurtful or of something contrary to many things one believes, it may take time for the belief to "sink in." cf. Valaris (2018)

[19] Is this to say that the logical and (causal) psychological aspects of reasoning relate to each other as different explanatory levels? I don't think so. The way I see it is that we are in the business of drawing

Similarly, there aren't mental processes that are necessarily involved in reasoning. Whether mental processes are involved, and which mental processes are involved, doesn't determine whether a particular thought or episode of thought is an instance of reasoning. A person can judge that *p as following from q* instantaneously, as if she is "just seeing" the connection. Or she can first imagine that *q* is true but *p* isn't; she might need to do some calculations; remember certain situations or conditionals; she might even need to write down the different possibilities, or speak to someone about it, before being able to judge that *p as following from q*. What makes a specific thought or an episode of thought an instance of reasoning is the involvement of judgments of this form, not the contribution of this or that mental process.

One might wonder whether reasoning doesn't require, perhaps not one particular mental process, but *some* mental process to be at work. Doesn't making a judgment depend on the functioning of psychological and neurological processes? Certainly so, but so too does believing something, or being in any other kind of mental state. That is, having any thought or attitude at all depends on the workings of neurological processes. Hence, on the neural level, every mental item can be seen as a process. The consequence of this is that, on this reading, calling reasoning a mental process doesn't do any work, at least not in distinguishing it from other items in the mental realm that we, on a mental, folk psychological level, call states or attitudes. Thus, the involvement of neurological processes doesn't imply that reasoning, on a mental, folk-psychological level, should be characterized as a mental process. These processes are so-called enabling conditions.

Another point of criticism might concern the following. How is it possible to think of reasoning as an *activity* of the person without it necessarily involving any mental process? This is a question meriting much broader treatment than I can give in this paper. For now, let me just mention that *process* is not the only form of activity in the mental realm. Judgment, for instance, is often categorized as a mental *act*. One main reason why such an act isn't a process is that it doesn't take time (Geach 1957; Roessler 2013a; Soteriou 2009). There is, for instance, no stopping halfway when one judges that *p as following from q*. Still, the person who makes the judgment can be considered to be active, because making the judgment depends on *her* taking it to be true: there is a form of agency, as Boyle (2011, 32) writes, 'whose exercise [does] not consist in actively changing things to produce a certain result,

different kind of connections (i.e., logical versus causal) in the world. But given that we ourselves have the capacity to draw the logical connections, these logical connections can (and should) inform the mental attitudes that we have. Thanks to Katrien Schaubroeck for pressing me on this point.

but in actively being a certain way.' In a similar vein, reasoning might be considered an activity, even if it isn't categorized as a process.

To further explicate the commitments of the form view regarding the relation between reasoning and a change in attitudes in particular, and between reasoning and mental processes in general, let me contrast it to Valaris' view (2018) that he has recently put forward. Valaris has argued that we should make a distinction between *deduction** and *reasoning**. As Valaris mentions, his 'use of these labels is to some extent stipulative' (Valaris 2018, fn. 1). I have marked them to distinguish them from the use of these notions in the rest of the paper. Deduction*, says Valaris, is a mental process of working out what follows from what. More specifically, deduction* is a mental process because it involves different steps: by eliminating different possibilities, one works out what follows from what, and comes to know a conditional. The result of such a deduction* is a 'doxastic attitude with conditional force, expressible as "A, given Γ"' (Valaris 2018, 11). Hence, it is a process with a specific result.

By contrast, reasoning* isn't a mental process, but does involve a change in view. The change in view involved in reasoning* is the adoption of a belief by taking it to follow from your premises – i.e., you adopt a belief with content of the form *p, given q*. The relation between deduction* and reasoning* is that reasoning* requires the recognition of a conditional. Sometimes such recognition results from deduction*, but it might also be the result of "just seeing" or, as Valaris strikingly adds, a bump on the head (which might make us vigilant about the nature of his account, I think).

What is striking in comparing Valaris' view with the form view is that both recognize the importance of conditionals. Where the form view uses the form of judgment that *p as following form q*, Valaris holds that deduction* involves a doxastic attitude with the form *A, given Γ* and that reasoning* involves a belief of the form *p, given q*. Given that *p as following from q* is equivalent with *p, given q*, the form of the attitudes under consideration is identical.

Still, Valaris' view differs from the form view in important ways. Valaris keeps both mental process and change in view in his account of reasoning (my terminology) but thinks that they belong to different aspects of reasoning: namely to deduction* and reasoning*. As has been explicated, on the form view, neither mental processes nor change in view are essential to reasoning. Rather, we can categorize a thought or an episode of thought as reasoning if it has a specific logical form, namely the form that one judges that *p as following from q* (or analogously, that one judges that *p, given q*). This means that, on the form view, deduction* and reasoning* do not explicate different aspects of reasoning but are both instances of

reasoning. They are instances of reasoning because they both involve the form of the judgment that *p as following from q*. That is to say that the mental process involved in deduction* and the change in view involved in reasoning* are instances of reasoning by virtue of the involvement of this form of judgment.

For instance, if we take a closer look at Valaris' exposition of the notion of deduction*, we see that it presupposes the capacity to make these form of judgments rather than explains it. 'Deduction,' as Valaris (2018, 8) writes, 'involves using information contained in your premises to *eliminate* or *exclude* possibilities… [W]hat I have in mind is simply recognizing that certain possibilities are *inconsistent with* your premises.' Now, my question is whether we can *use information* contained in our premises and whether we can *recognize* inconsistencies without using a capacity to think that *p* on the basis of something else, i.e., to judge that *p as following from q*? Using information in my premises to eliminate possibilities seems to me precisely an instance of judging that *p as following from q* (namely: judging that x is impossible given certain information in my premise), rather than a step leading up to a 'doxastic attitude with conditional force, expressible as "A, given Γ"' (Valaris 2018, 11). Similarly, recognizing that certain possibilities are inconsistent with your premises presupposes being able to judge something as following from something else (namely: judging that possibility x cannot be true, given that your premises are true).[20] Hence, I think that analyzing reasoning requires looking at the form of judgment involved and that this form of judgment is involved both in deduction* and reasoning* as they are characterized by Valaris.

7. Conclusion

I hope to have shown that the attitude view of reasoning, which states that reasoning is (1) a mental process that involves (2) a change in attitudes, faces numerous difficulties. Attitude views face the problem that there are many instances of reasoning without a change in attitudes, and they face problems in trying to formulate sufficiency conditions. In this paper, I have focused on the

[20] Valaris' notion of reasoning* is defined as the adoption of a belief based on reasoning (my terminology). I doubt whether beliefs adopted on the basis of reasoning always have the form *p, given q*. Suppose a person already believes that *q*, and believes that *if q then p*, but only just now these beliefs become relevant. By reasoning, she now also adopts the belief that *p*. Surely, her realization of her belief that *if q then p* is manifested in a judgment that *p as following from q*, which is the basis for the adoption of the belief that *p*. But there is no need for her to adopt the belief that *if q then p*. She already believed that. Hence, I don't see that reasoning* necessarily involves the adoption of a conditional belief. You have to put a belief in a conditional to *use*.

former. By giving examples of reasoning without a change in view, viz. (i) sustaining belief in the conclusion, (ii) Knorpp's example of puzzle-solving, (iii) non-formation of belief, (iv) hypothetical reasoning, (v) reasoning with an incoherent premise, and (vi) interpersonal reasoning, I have argued that a change of view isn't necessary for reasoning. As such, I have rejected (2) of attitude views. Moreover, as the example of hypothetical reasoning has made clear, advocates of the view that reasoning involves a change in attitudes, have to make clear whether their view includes categorical attitudes or hypothetical attitudes as well.

The lesson to draw from this is that we need a different approach to reasoning, one that includes instances of reasoning with and without a change in attitudes. The alternative view that I have sketched is that when a person reasons she (1) makes use of conditionals, manifested in (2) a judgment of the form *p as following form q*. I thereby take an analytic Aristotelian approach to reasoning, which holds that there are certain "form concepts" that require an analysis in terms of form of judgment, i.e., their logical form, rather than in terms of specific features or properties, or in terms of necessary and sufficient conditions. A consequence of this view is that, contrary to Valaris' recent proposal, mental processes nor change in view are necessary for reasoning. Whether a person is drawing up an argument, solving a puzzle, trying to follow someone else's line of reasoning, deliberating about what to believe, she judges that something follows from something else. And whether a person sees a truth-connection immediately or needs some time to imagine all the different possibilities before seeing it, she judges that something follows from something else. Judging that *p as following from q* is thus the form that each instance of reasoning takes. But judging isn't a mental process and thus (1) of attitude views is rejected too.

CHAPTER FOUR

Transparent Emotions? A Critical Analysis of Moran's Transparency Claim[1]

Abstract

I critically analyze Richard Moran's account of knowing one's own emotions, which depends on the Transparency Claim (TC) for self-knowledge. Applied to knowing one's own beliefs, TC states that when one is asked "Do you believe *p*?", one can answer by referencing reasons for believing *p*. TC works for belief because one is justified in believing that one believes *p* if one can give reasons for why *p* is true. Emotions, however, are also conceptually related to concerns; they involve a response to something one cares about. As a consequence, acquiring self-knowledge of one's emotions requires knowledge of other mental attitudes, which falls outside the scope of TC. Hence, TC cannot be applied to emotions.

1. Introduction

Richard Moran has developed a prominent account of self-knowledge[2] that depends on the Transparency Claim (TC), which is the claim that '[w]hen asked "Do I believe

---

[1] This chapter has been published as: Kloosterboer, Naomi. 2015. "Transparent Emotions? A critical analysis of Moran's Transparency Claim." *Philosophical Explorations* 18 (2): 246-258. Special issue "Self-knowledge in perspective," guest edited by Fleur Jongepier and Derek Strijbos.

[2] Bear in mind that philosophical discussions of self-knowledge are mainly about first-person awareness of one's mental attitudes (see Moran 2001, 31-32) and not about self-knowledge as we are familiar with in everyday usage: self-knowledge as knowing who we are, what is important to oneself, one's character traits or one's deeper concerns. However, this latter form of self-knowledge will play a role in the argument later.

P?", I can answer this question by consideration of the reasons in favor of P itself' (Moran 2003, 405; See also Moran 2001, 62-3). What is transparent about this claim is that the question about my mental attitude is seen as a question about its content: in a way, I look beyond the attitude to what the attitude is about. In this paper, I will discuss TC and address some of its problems. Especially, I will take issue with Moran's claim that TC can be applied to all mental attitudes, including one's emotions (Moran 2001, 64-5; 2012, 214; 2004a, 471). Since Moran does not explicate this claim, I will investigate whether and how TC can be applied to emotions.

Crucial to Moran's account of self-knowledge is his distinction between two different stances we can take toward our mental lives: a theoretical and a deliberative stance. Put concisely, from a theoretical perspective, I answer a question about whether I have a particular mental attitude by looking for evidence for my having the attitude. From a deliberative perspective, by contrast, I answer such a question in the way described by TC, namely by deliberating about the reasons in favor of or against the content of the attitude. Moran's account makes clear in what sense we are not mere bystanders of what goes on in our heads (theoretical stance), but that acquiring knowledge of our mental attitudes is related to how we see the world (deliberative stance). The following motivates this idea: although it is not unusual to say about someone else that she believes something that is actually false, from a first-person point of view, it does not make sense to say "*p* but I don't believe *p*". This exhibits a paradox or even irrationality, because believing *p* implies believing it to be true (see Moore 1993; and for the claim of irrationality, see Shah and Velleman 2005). Believing *p* and taking *p* to be true are conceptually connected. In general, from a first-person perspective, there is a conceptual relation between having an attitude with content C and endorsing C (or other forms of approval/disapproval, e.g. in the case of disbelief, rejecting C). As soon as one neglects the reasons one has pertaining to the content of the attitude, one distances oneself from one's relation to the world as a rational agent. This manifests an alienation of one's first-personal agential perspective and a failure of taking responsibility for being a rational agent. For this reason, Moran claims that the theoretical stance cannot account for the first-person character of self-knowledge nor for its relation to rational agency.

It is important to note that there are several issues in relation to TC that I will not address in this paper. For instance, I will not question whether Moran's account is a successful theory of self-knowledge of our beliefs or of first-person authority

(see Shoemaker 2003; McGeer 2007). Nor will I address the question whether TC can be applied to all instances of belief, for example, biased belief, dispositional beliefs, so-called hinge beliefs, and more (see Heal 2004; Cassam 2011). Furthermore, I will not criticize Moran's account on empirical grounds. That is, I will not examine whether we in fact acquire self-knowledge of our mental attitudes in the spirit of TC. My main aim is to show that TC cannot be applied to emotions. Therefore, for the purposes of this paper, I will assume that Moran's account applies to beliefs. If making up our mind about what to believe is a way of acquiring self-knowledge, how might this work in the case of emotion?

The paper is structured as follows. TC can be interpreted in different ways and therefore I will first discuss TC in more detail and propose what I take to be the most plausible interpretation (section 2). After addressing the question what relates TC to emotions in the first place (section 3), I will subsequently argue that it is impossible to apply TC to emotions (section 4). Finally, following the case of emotion, I will sketch some consequences for Moran's deliberative account of self-knowledge (section 5).

2. Moran's account of self-knowledge

Moran claims that a person comes to know her mental life by making up her mind about specific mental attitudes. In the case of belief, this can be explicated as follows. When someone asks me whether I currently believe something – for instance, whether it is raining – I do not introspectively observe my mental states to find evidence for this belief. Rather, I just look out into the world to see whether there are signs of rain. Based on my considerations about the weather, I then make up my mind and affirm the belief that it is raining. The idea that this form of deliberation results in self-knowledge of the mental attitude is known as TC. In this section, I will discuss TC in more detail. What does it amount to? And what makes it a legitimate claim?

Let us first take a closer look at what kind of claim Moran has in mind when talking about TC. Moran (2001, 62-3) discusses this question explicitly and makes clear that TC is not supposed to be an empirical claim. However, immediately after explaining that transparency is not something that is guaranteed empirically (or logically, but I will come back to that later), Moran writes that:

> With respect to believe, the claim of transparency is that within the first-person perspective, I treat the question of my belief about P as equivalent to the question of the truth of P. (Moran 2001, 63)

But a claim about how I treat a question is an empirical claim, so it remains unclear what kind of claim TC is, especially because Moran's formulations of TC range from an empirical claim to a capability claim to a normative ideal, and it seems as if Moran uses them interchangeably (see, for example, Moran 2001, 60; 2012, 212). Moreover, the sentence that follows the above indicates another ambiguity in Moran's account, namely that what is not empirically guaranteed is not TC itself but the relation between the two questions that TC is about. 'What I think we can see now,' Moran writes, '[. . .] is that the basis for this equivalence [i.e. between the two questions] hinges on the role of deliberative considerations about one's attitudes.' In other words, the relation exists in virtue of the fact 'that I address myself to the question of my state of mind in a deliberative spirit' (Moran 2001, 63). TC thus depends upon the stance I take vis-à-vis my own mental states. But then what does this say about the status of TC itself?

Without a canonical formulation of TC, I propose to understand TC as a claim about the capacity to answer a question about our mental attitudes in a certain way, that is, 'that a person *can* answer a question about her own belief by addressing herself to the corresponding question about the topic of that very belief' (Moran 2012, 212, my italics). This seems to be in line with Moran's overall account. Also, it seems to be the most plausible interpretation, given that the empirical interpretation and the normatively ideal interpretation of TC are seriously contested: there are many examples that undermine the claim that people always or normally answer questions about their mental life in accord with TC and also against the claim that it is always normatively ideal to do so (see Cassam 2011).[3]

The interpretation of TC as a capacity-claim means that it is a claim about how a person can make a certain transition between two questions. The first question is what Moran calls a self-related question about one's own mental attitude – "Do I believe that *p*?" – and the second question is what Moran calls a world-related question about the content of that attitude – "Is *p* the case?" or "Is *p* true?" But the labels self-related and world-related are not fully appropriate because it is unclear what to make of a question about a mental attitude about oneself, for example, "Do I believe that I am self-confident?" Which label should be applied to it? I suggest we circumvent this problem if we call the self-related question the attitude-question Qa

[3] See also Chapter 1.

and the world-related question the content-question Qc.[4] Thus, one question is about a particular mental attitude and the other about the content of that attitude, but their relation can be spelled out in more detail. It seems the case that Qa, that is, the question about a person S's mental attitudes, stands in a transparency relation to the relevant Qc about the content of that mental attitude, if Qa is answered by referencing the reasons that would justify an answer to the corresponding Qc (see Moran 2001, 61-2). The transition S can make between the questions is then that S answers Qa by answering Qc.

But this seems paradoxical: why is it legitimate to answer two logically distinct questions, that is, questions with different truth-conditions, by appealing to the same reasons? Why is it legitimate to arrive at a conclusion about whether I have the belief that *p* by considering whether *p* is the case? (see Moran 2012, 213; 2003, 404; 2004a, 466). This can be explained by an appeal to rational agency. In Moran's words:

> It would not . . . make sense to answer a question about my state of mind (e.g., my belief about the weather) by attending to a logically independent matter (the weather itself) unless it were legitimate for me to see myself as playing a role in the determination of what I believe generally,...in the sense that...the responsiveness to reasons that belongs to beliefs is an expression of the person's rational agency. (Moran 2012, 213)

According to Moran, TC presupposes rational agency in two ways. First of all, I need to see myself as playing a role and to take responsibility for this role. For Moran, taking responsibility means that I see it as up to me to avow myself on the matter. An avowal consists of a report of one's mental attitude including an explicit endorsement of its content. This is not to say that whether I believe something is wholly up to me – there is no voluntaristic implication. All it means is that I need to take responsibility for being reason-responsive. When asked whether I believe that *p*, I take responsibility if I follow the reasons I consider myself to have, if I arrive at my own conclusion by my deliberation on the matter. And as a consequence, avowing the belief that *p* expresses my endorsement of *p*. Moreover, it involves a commitment to the truth of *p*. As soon as I start doubting *p*'s truth or as soon as I reconsider the issue, the avowal ceases to exist (Moran 2001, 74-7, 80-2). Put

[4] In Chapter 1, Qa and Qc are characterized for belief as QB (question about the belief) and QP (question about the proposition), respectively.

differently, being committed to a certain mental attitude means making up your mind and sticking with it. Taking this responsibility is what Moran's envisages as taking a deliberative stance toward our mental life.

However, making up my mind in this way only succeeds if my belief actually is reason-responsive. The idea that a deliberative conclusion (or resolution as Moran calls it) constitutes my belief that *p* can only be true if we assume that 'my belief about...[*p*] is determined by my sense of the reasons in favor of [*p*], and not by forces independent of those reasons' (Moran 2012, 231). Accordingly, TC also presupposes that my mental attitudes are reason-responsive: I can exercise agency – be "active" or see myself as playing a role – with respect to my attitudes insofar as they are answerable to my sense of reasons and justification.[5] But what does it mean to say that mental attitudes are responsive to reasons? If this implies that they are formed through explicit deliberation, then many of our beliefs and other attitudes would not count as reason-responsive, since it is commonly accepted that many of our attitudes are not the result of deliberation but of unconscious processes. On a narrower understanding of reason-responsiveness, however, being aware of one's reasons for the attitudes one holds is not a necessary requirement. The point is rather that one should not be aware of a defeater for one's attitude.[6] In this way, attitudes that are not formed through explicit deliberation but are the result of unconscious processes can still be reason-responsive in the relevant sense, that is, if they are not in contradiction with the reasons one takes oneself to have. Being reason-responsive then means, in a minimal sense, that my attitudes are affected if I become aware of the presence of a defeater in the landscape of reasons. And being a full-blooded rational agent implies that I recognize and take responsibility for this relation between my mental attitudes and the reasons I consider myself to have in favor of or against those attitudes.

To conclude, TC is a claim about our capacity to make a transition between an attitude-related question (Qa) and a content-related question (Qc) that stand in a transparent relation to one another. This relation of transparency concerns the way in which these questions can be answered and the reasons that are taken into account, namely that I can answer Qa by answering the relevant Qc. But this can only be true insofar as my mental attitude is reason-sensitive, that is, is actually

---

[5] For a helpful elucidation of what kind of agency is involved, see Boyle (2011) and Hieronymi (2009). Hieronymi dubs the agency exercised in being reason-responsive "evaluative control".

[6] This is also Shoemaker's interpretation (2003, 396).

determined by my answer to Qc. This means that I can answer Qa by considering the relevant Qc if I take a deliberative stance vis-à-vis my mental attitudes, a stance from which 'I take myself to be responsible for making my belief conform to my sense of the reasons in favor or against,' and consequently, that if I avow the belief that *p*, I thereby express my commitment to its truth (Moran 2003, 406). TC thus boils down to a claim about a person's agential capacity to determine her state of mind by making up her mind.

3. Emotions as intentional attitudes

Although Moran explicates his account only for belief, he asserts that TC is also applicable to emotions and other mental attitudes. One might wonder, however, why emotions are candidates for TC in the first place. Most of the time, emotions seem to befall us without any deliberative activity on our part. Do we ever make up our minds about our emotions?

From an empirical perspective, it is difficult to say how often or seldom we do this. However, it seems plausible that we make up our minds about our emotions when we do not know what we feel or what we should feel. And maybe such situations are not that uncommon: if we consider music, literature and art, the ambiguity and uncertainty of our emotional lives stand out. Not to mention psychology and psychiatry, where thinking and talking about how we feel are a crucial part of therapy. However, as said in the introduction, this paper concerns the question whether applying TC to emotion is possible, that is, whether we have the ability to answer Qa about emotions in accord with TC. Therefore, we should ask: does it make sense to apply TC to emotions? Before addressing this question, two qualifications are in order.

First, recall that TC is about mental attitudes that have intentional content C about which one can ask a Qc, that is, a question about the justification of C. This means that it makes sense to apply TC to mental attitudes that have such content. Do emotions have such content? I will not argue for a specific theory of emotions here, but only assume that, in the field of feelings, emotions, moods and so on, some of them have intentional content.[7] For instance, if you are angry, you are often angry with someone, at someone's action, or at something or a situation.

[7] For the stronger claim that for an attitude to be an emotion, it must be intentional, see (Moran 2001, 54; De Sousa 2007; Teroni 2007; Döring 2007). For the claim that even feelings are intentional, see, for example, Goldie (2002).

Second, emotions put these circumstances in a certain light: your anger says something related to the thing you are angry with or at, for example, that it is offensive, irritating or hurtful. As Moran puts it: emotions are intentional attitudes that express a person's "understanding of the world", "way of seeing the world", or are part of that person's "total orientation", or "total outlook" (Moran 2001, 41-2, 50-1). Put differently, emotions are part of one's evaluative perspective (see Helm 2010, 315).[8] If an emotion entails an evaluation of the world, we can ask whether the evaluation is justified, whether the emotion purports a right way to see the world, and we can ask someone for her reasons why she sees the world in this way. Accordingly, these emotions are related to our sense of reasons and justification. In what follows, I will only be concerned with those emotional attitudes that have an intentional object and entail an evaluation of that object, which can be subjected to question(s) of justification (i.e. to Qc).

4. Transparency and emotions

To find out whether TC applies to emotions, we need to discern Qa and Qc for emotions. In this way, the differences between beliefs and emotions will become apparent and, as a result, we will be able to determine whether TC can be applied to emotions.[9]

Even if Moran has not spelled out TC for emotions, his work suggests a particular Qa and Qc for emotions. Consider, for instance, the case described by Moran of a patient undergoing psychoanalysis, who cannot acquire self-knowledge in the way described by TC:

[8] This is not to say that emotions are as reason-responsive as our beliefs. Emotions are known for their impenetrability (see Döring 2007). But we still criticize them if they diverge with one's evaluative perspective (see Smith 2005).

[9] An interpretation of TC that can be applied to all mental attitudes is proposed by Finkelstein (2012, 103), and endorsed by Cassam (2014, 4): 'The question of whether I believe that P is, for me, transparent to the question of what I ought rationally to believe – i.e. to the question of whether the reasons require me to believe that P. I can answer the former question by answering the latter.' Finkelstein's formulation of Qc is "Ought I rationally believe that P?" in which, so the thought goes, believe can be substituted by desire, feel, intend and so on. Although it seems to be a very elegant solution to the problem of applying TC to other attitudes than belief, I think it is incorrect. The main reason why it does not seem right to me is that it is not a question about the content of the mental attitude, but another inward-directed or self-related question, namely about what kind of believing it is, that is, whether it is rational. Clearly, more needs to be said about this, but that will have to wait for another time.

> In various familiar therapeutic contexts, for instance, the manner in which the analysand [subject of psychoanalysis] becomes aware of various of her beliefs and other attitudes does not necessarily conform to the Transparency Condition. 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 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 the attitude. [. . .] [S]he will not in her present state affirm the judgment that this person has in fact betrayed her. When the belief is described, it is kept within brackets of the psychological operator, 'believe'; that is, she will affirm the psychological judgment "I believe that P," but will not avow the embedded proposition P itself.[10] (Moran 2001, 85)

Moran seems to say here that the Qa and the relevant Qc of the emotion are:

> Qa: Do I feel betrayed by this person?
> Qc: Did this person in fact betray me?

The formulation of the Qa of emotion "Do I feel betrayed by this person?" seems right to me. I propose the following general form:

> Qa*: Do I have emotion X about O (or P)? (Or Qa*: Do I feel X at O (or P)?)

In this way, we can fill in the kind of emotion for X and then for either O or P the object, person or proposition the emotion is directed at. Accordingly, we should ask the analysand "Do you feel betrayed by this person?" or "Do you feel betrayed by what this person has done?"

---

[10] Shoemaker has criticized this last sentence (2003, 397), and Moran has corrected it (2003, 410). It should actually say: she will avow the psychological judgment "I believe that *p*" but will not affirm the embedded proposition of *p* itself.

Let us now examine the Qc of emotion. If we go back to Moran's example of the analysand, we see that there is an important change from an emotion to a belief: it is about someone who feels betrayed but cannot avow the belief that she has been betrayed. Accordingly, for Moran, the Qc seems to be: "Is it true that this person betrayed you?" which is an instance of the Qc of belief: "Is *p* true?" But how can this question be specifically about the emotion of feeling betrayed and not merely about whether one believes that one is betrayed? Moran seems to assume that the reasons that are relevant in deliberating about one's beliefs are the same reasons that one would address in deliberating about one's emotions. I will argue, however, that Moran neglects a difference between beliefs and emotions, which affects the way TC can be applied to emotion.

As we have seen, the reasons that are relevant in deliberating about beliefs have to do with the truth-value of the content of the belief. Determining whether to believe something comes down to determining whether the belief is true or not, because beliefs aim at truth (Moran 2001, 52, 69-77; Shah and Velleman 2005, 498). Only those considerations that are about the content of the belief play a role in making up my mind about it. And that beliefs are related to truth in this way is precisely the reason why TC is applicable to them: if I deliberate about the Qa "Do I believe that *p*?", I turn myself to the Qc "Is *p* true?" and I only regard content-related considerations of that particular belief.

But emotions stand in a different relation to reasons. For example, I am not angry with a friend for forgetting our date only because it is the case that she forgot our date. Of course, my anger is only justified if she has in fact forgotten the appointment (though my anger still makes sense if I only believe this to be the case). However, if I want to deliberate about whether I am angry with my friend or not, saying this is true does nothing to explain why being angry is the right way to feel. Emotions do not aim at truth in the way beliefs do (see De Sousa 2007, 328; Teroni 2007, 399). Hence, the Qc of emotion cannot be "Is *p* true?" We need to consider other reasons to make up our mind about our emotions. But what reasons are those? Can we articulate an alternative Qc of emotion?

Recall that emotions entail a way of seeing the world: becoming angry puts the situation in an evaluative light; for example, that the object concerned was offensive. Accordingly, my anger seems to be justified if it purports an apt evaluation of the situation, that is, if forgetting a date is the right sort of thing to be evaluated as offensive (see Teroni 2007, 404). In giving reasons for my anger, I need to say something about why forgetting a date is something to be angry about; I need to give an evaluation of it that makes my anger an apt response (see Smith 2005, 250-3). The relevant Qc could thus be: "Does the fact that my friend forgot our appointment

have features that make anger an appropriate response?" More generally, the Qc of emotion could be formulated as follows:

> Qc*: Does P (or O) have features that make X an appropriate response?

So in the case of Moran's example of the "analysand", if she wants to answer the question "Do I feel betrayed by what this person has done?" in the spirit of TC, she should address the question "Does what this person has done have features that make feeling betrayed an appropriate response?" It seems therefore as if Qc* is a good candidate for a Qc of emotion.

However, there remains a problem with this formulation of the Qc. The point is that we need a formulation with which we can determine whether the emotion is justified, and not merely whether the evaluation inherent in the emotion is apt. Otherwise, there would be no difference between a mere evaluation and an emotion. And there is at least this minimal yet very important difference, namely, that we make value judgments about anything, but we only feel an emotion if something matters to us (see Helm 2010). Our emotions are conceptually related to our concerns in the sense that they are responses to things that are of our concern. And the features of "what this person has done" simply fail to account for this. In order to explain this, I need to explicate the conceptual relation between emotions and concerns in more detail.

On the one hand, to be concerned about something is constituted by having certain emotions. In the words of Bennett Helm, '. . . it is hard to make sense of someone as caring about something if he or she does not respond emotionally no matter what when it is affected favorably or adversely' (2010, 311). Nomy Arpaly also neatly describes how emotions are one of the constitutive elements of what it is to be concerned about someone or something:

> Other things being equal, caring about a team makes wins pleasant and losses painful. More than this, the person who cares about a team is likely to experience shame at its bad performance, pride at its good performance, anxiety when an important game approaches, a sense of utter despair if it turns out that a key player has been involved in a serious drug fraud, and other such emotions that utterly baffle the person who does not possess such a concern. (Arpaly 2003, 86)

If one does not respond emotionally when something that one cares about is affected, one cannot be said to genuinely care about it: to be concerned about something requires that the concern resonates through one's emotions.

On the other hand, this implies that if one responds emotionally to something, it should be important to one. If a person is not concerned about her safety, why feel fear? If not concerned about someone's well-being, why feel bad if something bad befalls that person? If not concerned about the team in Arpaly's example, why feel all those emotions if something good or bad happens? What these examples show is that, conceptually, having an emotion expresses a commitment to something or someone that matters to the person in question. Importantly, this implies that the justification of emotions, unlike beliefs, also depends on the question whether the emotion is a response to something or someone that is of one's concern. Emotions should not only be sensitive to content-related reasons, that is, reasons that allow one to determine whether the evaluation is apt, but also to reasons related to what is important to one: the justification of emotions depends on the question why something is dangerous, hurtful, offensive, joyful, shameful or thrilling, *for the person in question* (see Helm 2009, 250-1). To put it differently, emotions should not only be correctly tuned to the world; they also involve an appropriate attunement to things that concern us.

The upshot of the conceptual relation between emotions and concerns is then that deliberation about what to feel cannot be limited to reflection on facts relevant to the specific evaluative content of the emotion but includes considerations about what is important to someone. To determine whether I have a certain emotion, I also need to answer, apart from the relevant Qc, another question, namely

Qa-care: Do I care about P (or O)?

The problem for TC is that this latter question is another attitude-question and not a content-question. Hence, TC fails in the case of emotion because the justification of the emotional attitude not only depends on what the emotion is about, but also on how it relates to one's other mental attitudes.

However, one might think of two immediate objections. The first objection is that the supplement "for the person in question" can be included in the Qc. Why can the Qc not be: "Does O (or P) have features that make X an appropriate response for me?" But the supplement "for me" actually changes the question. Instead of being a question about the aptness of the evaluation in light of the situation, it is now a question about the aptness of the evaluation in light of the person I am. And this is a question that can only be answered by referring to other mental attitudes of mine.

Telling whether something is hurtful, offensive, or joyful for a specific person is grounded in considerations that depend upon who that person is, with certain character traits, concerns, plans, ambitions, fears, vulnerabilities, relations to other persons and so on.

The second objection is that emotions should not be tuned to our actual concerns, but to what should concern us. And maybe what should concern us can be determined without appealing to other mental attitudes, that is, only by considering reasons in favor of or against the appropriateness of the object of one's concern. Naturally, it is very important that one does not just care about anything, but about the relevant things. We want to be a person who cares about the things she should care about (see Jones 2004, 343). Therefore, the question whether something is an appropriate object of concern is very important in determining whether to care about something. Nevertheless, these considerations remain inconclusive with respect to which of all those appropriate things I actually care or can come to care about. Here again, we are back at the conceptual relation between concerns and emotions; a relation that extends to other mental attitudes and to actions as well. According to Helm, caring means that the object of one's care is the focus of a pattern of emotions, desires, judgments, intentions and actions (2010, 311-5). Similarly, Arpaly develops an account in which three features constitute concerns: a motivational, emotional and cognitive one (2003, 85-7). These features determine the strength of your concern: the more it takes to stop you from acting out of your concern, the more and stronger emotions you have concerning the object of your concern, and the more attuned you are to notice circumstances of and about the object of your concern, the stronger your concern is (see also Smith 2005). Now, for a concern being constituted by these other attitudes and actions means that one can only be said to genuinely care about something if one has these other relevant attitudes and performs relevant actions. Consequently, deliberating about whether something or someone is an appropriate object of care does not answer the question whether it is an appropriate object of care for you: one needs to take into account who one is, who one wants to be and who one can become to determine whether or not one cares or can come to care about something.

Let us return to Moran's example of the "analysand". According to Moran, what goes wrong in her case is that she cannot endorse facts about the person who betrayed her:

> It is because her awareness of her sense of betrayal is detached from her sense of the reasons, if any, supporting it that she cannot become aware of it by reflecting on that very person, the one by whom she feels

> betrayed. The rationality of her response requires that she be in a position to avow her attitude toward him, and not just describe or report on it, however accurately, for it is only from the position of avowal that she is necessarily acknowledging facts about him as internally relevant to that attitude (say, as justifying or undermining it), and thereby (also) as relevant to the fully empirical question of whether it remains true that she indeed has this sense of being betrayed by him. Otherwise, her own sense of the truth about that person floats free of her sense of what sustains her attitude toward him. (Moran 2001, 93)

In making up her mind about feeling betrayed, the patient should reflect "on that very person", on "facts about him", and find out the truth about "that person". But, as we have seen, the problem might also be that she does not acknowledge facts about herself. She also needs to acknowledge why that person and those facts about him are important for her. And to do this, reflecting about what the other person has done will not help her forward. Instead, she needs to reflect on, for example, what he means to her, her expectations of the relationship, and her fears. Therefore, not only the patient's "sense of the truth" about that person is relevant, but also her sense of the truth about herself. The difference between being able to attribute the feeling of betrayal and to avow the feeling of betrayal lies as much in the acceptance of what that other person has done as in the acceptance of what is important to oneself and what kind of person one is.

To conclude, Moran's claim that TC can also be applied to emotions does not hold. As we have seen, reasons for having an emotion include not only content-related considerations but also considerations about other attitudes, for example, about what is important to one. This means that the reasons I have to answer the Qc* "Does P (or O) have features that make X an appropriate response?" do not provide an answer to the Qa* "Do I have emotion X about O (or P)?" Features of P (or O) are among the reasons that justify having the emotion, but based on those reasons alone, I cannot, without further considerations about what is important to me, determine whether I should have the emotion or not. I also need to answer Qa-care: "Do I care about P (or O)?" So, TC cannot be applied to emotions, because there is no deliberative question with which I can determine whether the emotion is justified that is exclusively directed at the content of the emotion, or can be answered by content-related considerations alone.

5. Commitment and agency

In the previous section, we have seen that it is problematic to apply TC to emotions. In this section, I sketch some consequences of the relation between emotions and concerns for Moran's account of commitment and agency. Moran's deliberative account of agency and commitment has already been criticized for being too rationalistic or idealized (see McGeer 2007; Cassam 2014). My criticism, however, does not issue from the fact that his account is deliberative, but from the fact that Moran neglects the conceptual relation between our mental attitudes and what concerns us. As a consequence, he misunderstands which considerations play a role in deliberation and thereby misunderstands some central features of what it means to be an agent, namely a person who is not only searching for the truth of things, but to whom the world matters in a specific way.

Let me explicate this more carefully by introducing the first consequence for Moran's account of reflective endorsement and commitment. According to Moran, in avowing an attitude, we thereby express a commitment to its truth and to what that requires of us, for example, not holding a contradictory attitude. If you deliberate about a certain Qc and come to a conclusion, this thereby constitutes a commitment. The state of being committed can therefore be understood as a state of having made up your mind about one specific attitude. The question is whether this notion of commitment can account for a genuine commitment. The normative relation between emotions and concerns highlights an alternative idea of what it means to be committed to something. If I avow a certain emotion or a concern, what do I commit myself to?

As stated above, being concerned about something or someone only makes sense if the object of one's concern is the focus of a pattern of actions, desires and emotions. Let me say more about this pattern in the case of emotions. It is actually an implication of the conceptual relation between emotions and concerns that emotions only make sense if they form patterns. As stated by Helm: '. . . to feel one emotion is to be rationally committed to feeling a whole pattern of other emotions with a common focus' (2009, 251). If I hope I will catch a train, I should also feel subsequent emotions if I either miss or catch the train, for example, feel disappointed or happy. This means that if I feel fear about a threat, I am committed to also hope that the threat will not actually be realized, to feel bad if it does happen, and to feel relieved when it does not. In other words, the state of being committed to something depends more upon having the relevant patterns of emotions, desires and actions than on reflectively endorsing the content of the attitude. Again, Arpaly is on the right track in saying that:

> Two people can reflectively endorse identical things but be very different in their level of concern for these things. Erica and I may reflectively endorse the same kind of political action, but she may be more concerned with it than I am, which may explain why she is at a demonstration while I am writing. It is also natural to say that I am less committed to political action than Erica is . . . [W]e may deeply care about things that we do not reflectively endorse at all. Tamara may care deeply about Todd and her relationship with him even though she believes she should not do so, or even though she is utterly unaware of her deep concern, ignoring it in practical deliberation. (Arpaly 2003, 85)

This means that if I declare myself to be committed to, for example, being on time, I also have to worry about being on time, feel ashamed if I am not, excuse myself for not being on time, and take precautions to be on time. If I were a person for whom none of these things matter, the fact that in deliberating I decide to be a conscientious person will have little effect. A genuine commitment, therefore, requires that we have the right emotional responses to the appropriate things and make efforts to avoid and achieve the right things. Being committed to something implies that one is concerned about it, which can only be genuine if the commitment resonates through one's emotions, desires, intentions and actions.

A further implication of this might be that Moran's account of self-knowledge, that is, self-awareness of mental attitudes, presupposes another form of self-knowledge, namely knowing what kind of person one is. This is a bold claim, so I want to stress that this is a suggestion. However, there is a reason to take up this line of thought. For instance, if I sincerely declare that I want to be on time, that is, if I commit myself to being on time, and I do not worry about being on time, nor feel sorry if I am late, nor take precautions to be on time, then this shows that either I did not care about being on time at all or that I did not care about it enough. My declaration then conflicts with who I am and not with whether being on time is an appropriate object of concern. In other words, sincerely declaring something that is not at all (or not enough) related to what is important to me shows that I apparently do not know what kind of person I am. My point is that only avowals that express who I am or who I can become can be seen as a genuine commitment. And this assumes that I have the capacity to relate my current attitudes to what is important to myself. For I cannot know what is important to me by means of deliberating about content-related considerations of a specific attitude alone precisely because

importance and commitment are conceptually linked to already having or coming to have the relevant attitudes that express them.

In addition, emotions are not the only attitudes that invoke the "concern"-critique of Moran's view. All attitudes that are related to one's evaluative perspective exhibit a conceptual relation to concerns. Belief and other mental attitudes that aim at truth seem to be the exception rather than the rule. So, analogous to the case of emotions, having a desire does not make sense without the thing desired being related to something that is important to you, and intending to do something does not make sense if it would not matter to you whether you actually do it or not.

To conclude, one of Moran's key aims was to save the discussion about self-knowledge from its exclusive epistemic approach and to put agency at the core of the picture of self-knowledge. But it seems that Moran's conception of agency remains focused on epistemic agency, because Moran's agent only deliberates about the question whether her attitudes are true. He thereby neglects other considerations that are relevant to the practical rationality of the agent. To get epistemic and practical agency into view, making up our mind includes taking into account other mental attitudes we have and so paying attention to who we are. Something that, in Moran's account, implies alienating oneself from one's relation to the world as a rational agent. The criticism of Moran's conception of agency and commitment displayed here is only a sketch. The main point, however, should be clear: Moran's idea that rational agency consists in deliberating about the truth of our mental attitudes neglects important aspects of our practical agency, in particular that we are agents to whom things matter, with certain projects, relationships, vulnerabilities and peculiarities.

6. Conclusion

We have seen that TC is a claim about the agential ability of a person to make up her mind about a Qa – whether she has a mental attitude – by considering reasons that provide an answer to the relevant Qc – whether the attitude is justified. In this process of deliberation, the considerations taken into account are limited to those related to the truth-value of the content of the attitude and hence to the justification of the attitude. The justification of emotions, however, depends not only on the question whether the content of the emotion is true or apt, but also on the question whether the emotion is a response to something that is of one's concern. Our emotions only make sense if they are responses to things that are important to us.

So, in the case of emotions, considerations about what is important to one should also play a role in deliberation. However, in Moran's account, these considerations are excluded from the first-person deliberative stance, because they are not about the content of the specific emotion itself. Hence, TC cannot be applied to emotions.

This is not to overlook the importance of the relation between emotions, our evaluative outlook, and our sense of reasons. On the contrary, this relation is crucial in helping us to understand which emotion can be seen as rational, appropriate or as an expression of the person one is. Moreover, it furthers our understanding of those emotions (or lack thereof) that strike us as irrational or even pathological, such as phobias, inadequate affect in psychotic disorders and lack of remorse in sociopaths. It is even the case that an emotion might correct our sense of reasons. Sometimes we do not take ourselves to have a reason to do something in a situation where we do have such a reason: in such a case, our emotions might just point out what is important to us.

CHAPTER FIVE

The Status of Avowal in Substantial Self-Knowledge

Abstract

In recent discussions on self-knowledge of intentional mental attitudes, a distinction is made between so-called trivial and substantial self-knowledge, where a subject's self-knowledge is substantial if the object of knowledge is significant to her life and self-conception (cf. Cassam 2014; Schwitzgebel 2012). The distinction is used to argue against, among others, Richard Moran's account of self-knowledge (2001). It is claimed that substantial attitudes aren't revealed in what a person avows, but in her patterns of action and reaction. In this paper, I will argue to the contrary: even if such patterns of action and reaction form part of coming to know my substantial mental attitudes, avowing these attitudes remains essential and has a unique status in coming to know them. Avowal is essential to knowing one's substantial mental attitudes, I claim, because these attitudes require one to have a self-conception. It requires a person to take on the burden of agency.

1. Introduction

In recent discussions on self-knowledge of intentional mental attitudes, a distinction is made between so-called trivial and substantial self-knowledge, where a subject's self-knowledge is substantial if the object of knowledge is significant to her life and self-conception (cf. Cassam 2014; Schwitzgebel 2012).[1] The distinction is used to argue against, among others, Richard Moran's account of self-knowledge (2001). It is claimed that whereas perhaps trivial self-knowledge is a result of the special

[1] Although "trivial" and "substantial" might not be the best terms to depict these different kinds of self-knowledge – after all, what is coined "trivial" in this distinction is *also* crucial in our lives – I stick to the terminology for simplicity's sake.

relation a person has to her own mental life as portrayed by Moran – especially, that she is in a position to avow her mental attitudes – substantial self-knowledge is not. Even if my knowledge of, say, believing that it is raining or wanting to drink ginger beer (i.e., trivial attitudes) is a consequence of the special relation I have to these attitudes, my knowledge of substantial mental attitudes such as caring about my job or believing that women and men deserve equal treatment is not. If I were to seek knowledge of these kind of attitudes, so the suggestion is, I should observe and interpret the patterns of my actions and reactions, or even better, turn to my peers (cf. Cassam 2014; Schwitzgebel 2012).

In this paper, I want to argue to the contrary: even if such patterns of action and reaction form part of coming to know my substantial mental attitudes, avowing these attitudes remains necessary and has a unique status in coming to know them. Avowal is essential to knowing one's substantial mental attitudes, I claim, because knowing the engagement pertaining to these attitudes requires one to have a self-conception.

The paper starts with two preliminary sections. I will first address the notion of avowal and its relation to agency and self-knowledge (section 1). Next, I take a closer look at what the problem with the role of avowal in knowing one's substantial mental attitudes is (section 2). Then, I turn to the arguments. The first question to be addressed is whether a person can know her substantial mental attitudes without avowal (section 3). I will argue that mental attitudes cannot be "revealed" in patterns of action and reaction. The basic problem is that the significance of patterns of action and reaction, and what such patterns tell about our attitudes, ultimately depends on avowal. This also means that current popular accounts of self-knowledge that seek to relativize or even undermine the special relation a person has to her mental attitudes, which is expressed in avowing them, cannot succeed without recognizing a unique role for avowal.[2]

Finally, I will focus on what the positive role of avowal can be (section 4). Even if we accept the possibility and prevalence of a gap between a person's avowals and her patterns of action and reaction, it isn't thereby determined what such a gap implies. It is often assumed that such a gap implies ignorance, or that making mistakes demonstrates that the relevant capacity is unreliable and should be mistrusted.[3] What I will argue is that ignorance or unreliability doesn't necessarily follow from making mistakes; such mistakes might actually be inherent in the kind of process or the kind of capacity at issue. In this light, I will suggest that acquiring

[2] For such accounts, cf. Cassam (2014); Lawlor (2009); Schwitzgebel (2012); Wilson (2002).
[3] Cf. Carruthers (2011); Cassam (2014); Doris (2015); Schwitzgebel (2010), among others.

self-knowledge of substantial mental attitudes is a struggle, namely, to fulfill the commitments pertaining to these attitudes. It requires a person to take on the burden of agency – to take responsibility for who she is and putting herself at risk of being challenged and making mistakes. Or so I will argue.

2. Avowal, self-knowledge and agency

This first preliminary section concerns the role of avowal in self-knowledge and its relation to agency. Why would we even think that avowal has any role to play in achieving self-knowledge of our intentional mental attitudes? My point of departure to answer this question is Moran's (2001) influential account of self-knowledge.

For Moran, achieving self-knowledge is a matter of avowing one's mental attitude. An avowal consists of a self-attribution (either verbally, in inner speech or in thought) of one's mental attitude including an explicit endorsement of its content. In the case of belief, to avow my belief is to express my 'own present commitment to the truth of the proposition in question' (Moran 2001, 86). This also implies that the self-knowledge that Moran seeks to characterize is knowledge, not just of my having a belief, but of my commitment towards *p*. It is thus knowledge that puts me in a position 'to speak of [my] conviction about the facts' (Moran 2001, 76).[4]

That avowal plays such a crucial role in Moran's account is a consequence of the idea that my relation to my mental life is different from my relation to the mental life of others (or of other's relation to my mental life). This essential difference is, according to Moran, not a difference in epistemic access or privilege but a difference in the way a person is involved in what her mental life is, namely as mental agent. My relation to my mental attitudes is not that of an expert witness, or of a bystander who happens to have the best information about my mental attitudes. Different from being a witness or bystander, I do not merely register what is present in my mind. Rather, my mental attitudes express my relation to the world, my *stance* or *grasp* on how things stand in the world.[5] As such, they must be seen by me 'as expressive of

[4] Moran's view of avowal is often portrayed in a way that it necessarily involves considering reasons and determining whether it would be rational to have the attitude (cf. Cassam 2014; Finkelstein 2012). At some points, Moran seems to be committed to such a demanding picture of avowal, though another interpretation of Moran's view is also possible (see also Boyle 2015). In this paper, I will leave this issue aside and go with a less demanding notion of avowal. This notion captures, not the rational status of a person's attitudes, but her stance towards them, i.e., a stance that manifests the relation between a person's self-attribution and her view of the world at large (hence, of not being alienated from one's mental attitude, as Moran would say).

[5] Importantly, this claim holds for self-knowledge of *intentional mental attitudes*. These attitudes, such as beliefs, emotions, desires and intentions, are fundamentally different from sensations, headaches and

[my] various and evolving relations to [my] environment, and not as a mere succession of representations (to which, for some reason, [I am] the only witness)' (Moran 2001, 32).[6]

Importantly, respecting this relation between my attitudes and my 'evolving relations to the environment' involves taking responsibility for my mental attitudes: after all, they express *my own* view of things. This is not to say that these attitudes are under my voluntary control; rather, I take responsibility for them in the same way as I may take responsibility for the conclusion of my reasoning, or for the love I feel for someone, not because I could reason in whatever way I wish or love whomever I favor, but precisely because the reasoning and love are expressive of *my own stance* (cf. Moran 2008). It is often presumed that one exercises one's agential capacities only in forming or changing one's mental attitudes, but I take it that the agency involved in avowal is best understood if one sees these capacities at work also in *having* a mental attitude. We might say that a subject's agential capacities are at work insofar as her mental attitude is not a *given* fact, but something she must settle and sustain (cf. Moran 2001, 77; Boyle 2015, 341).[7]

The characterization of self-knowledge arrived at is one where one's knowledge of a mental attitude dovetails with being committed to the grasp of the world purported by the attitude in question. Arriving at this form of self-knowledge requires avowal. Through avowal, a person fulfills the condition that she sees her attitudes as expressive of her grasp on the world at large and the condition that she exercises her mental agency in having or taking a stance.

Finally, as a last remark, Moran's view is often portrayed as saying that avowal in and of itself is sufficient for self-knowledge. This seems to be implied in Moran's claim that avowal constitutes self-knowledge. But Moran also seems to hold

---

heart rates, because they involve, for the subject of those states, a characteristic grasp of the world. That is to say that these attitudes involve, from a first-person perspective, grasping the (propositional) object of those states *as true*, *as to be done*, *as dangerous*, etcetera.

[6] This is reflected in Moran's notion of the *deliberative stance*, which is to be distinguished from the *theoretical stance*. These stances correspond with two kinds of questions about and inquiries into one's mental life. A theoretical question about one's mental life is 'one that is answered by discovery of the fact of which one was ignorant', Moran explains, 'whereas a practical or deliberative question is answered by a decision or commitment of some sort, and it is not a response to ignorance of some antecedent fact about oneself' (2001, 58).

[7] This form of mental agency is further developed in Boyle (2011); Hieronymi (2009); and Moran (2012). Moran also extensively focuses on deliberation and (justifying) reasons. I leave this aspect of Moran aside and focus on commitments and the agency inherent in *being committed*. What is thus left open by my discussion is whether and how deliberation and justification ought to be involved in an account of self-knowledge. This paper thus isn't a *defense* of Moran's account. Rather, it investigates the role of avowal in substantial self-knowledge.

that avowal isn't in and of itself a sufficient condition.[8] Rather than trying to answer this exegetical question, this paper focuses on avowal as a necessary condition of self-knowledge and on its unique status in acquiring self-knowledge. Consequently, this paper isn't directly concerned with the questions whether and how avowal constitutes self-knowledge.

3. Substantial self-knowledge and care

In order to address the role of avowal in acquiring substantial self-knowledge, I will focus on *care* as substantial mental attitude. Care can mean different things. You can be a caring person (character trait), or you can take care of the bills (action), or you can care about your job (intentional attitude). It is this latter notion that's implied in the discussion. I focus on care for the reason that it is uncontroversial regarding its 'attitudinal' nature (as opposed to, e.g., character traits), but especially because most substantial attitudes will involve care: if attitudes are to be significant to one's life, then this will presumably involve that the object of these attitudes are related to one's cares.[9]

What is it to care about someone or something? What commitments are involved in caring? A basic commitment inherent in caring about X seems to be that X is important for me. But if X is important for me, I am also committed to integrate X in my life in relevant ways. Hence, if one cares about X, one is committed to a whole pattern of other mental attitudes and actions. For instance, if I say that I care about my job, this seems to be sincere only if I also want to do my job well, would regret missing an important meeting, make sure to put effort in my work and enjoy doing my job. That is to say, caring about my job requires the right kind of *engagement* on my part.[10]

---

[8] In his response to Shoemaker and O'Brien, for instance, Moran (2003, 20) emphasizes how the capacity to make up one's mind depends on prior experience and evidence. He asserts that whereas an adequate history and the right empirical facts must be *in place* to make up one's mind, it cannot *replace* making up one's mind: '…there is a great deal of empirical complexity that must be assumed and relied on for something as simple as ordering from a menu, and when all this is in place, the transition from not knowing to knowing what one will have is made by arriving at a decision'.

[9] This is also reflected in Cassam's (2014, 31-2) value-condition, which is part of his definition of substantial self-knowledge.

[10] Cf. Arpaly (2003); Helm (2010); Smith (2005); Seidman (2016). In Helm's terminology, caring about X means that X is the focus of a pattern of emotions, desires, judgments, intentions and actions (2010, 311-5). According to Arpaly, caring is constituted by three types of engagement: a motivational, emotional and cognitive one (2003, 85-7).

The right kind of engagement, i.e., the right pattern of actions and reactions, involves both a requirement of coherency and a requirement of robustness. The engagement should be, first, reasonably coherent considering a range of actions and reactions that I may have at the same time. If I am meticulous about finishing a report before the deadline, put an enormous amount of effort in it, but feel no relief when the job is done, nor pay attention to how well the report is received, something seems to be off. Secondly, the engagement should also normally be robust. First, it should normally exist for an extended amount of time: caring about my job is not something I can do for just a day.[11] Secondly, there also seem to be constraints on how often or in what way a substantial attitude changes: normally, it shouldn't change randomly, nor very often.

What follows from the characterization of care as involving a particular engagement (patterns of action and reaction) is that care isn't only manifested in what a person says she is committed to, but also in her actions and reactions. This means there is reason to doubt the unique status of avowal. Suppose I avow my belief that taking care of the environment is very important. I thereby express my commitment to taking it to be true that taking care of the environment is very important. But suppose also that I don't recycle trash, I travel by airplane even if going by train is a viable alternative, I leave the lights on, etcetera. If I don't act on my professed belief, then it surely seems as if I don't *really* think that taking care of the environment is very important. If a person's avowal of a substantial attitude doesn't resonate in her actions and reactions, then what significance does her avowal have?

Critics of the unique status of avowal with respect to substantial self-knowledge, seem to claim that the role of avowal in achieving self-knowledge is only marginal. In this view, substantial attitudes aren't revealed in what a person avows, but in her patterns of action and reaction. Patterns that someone else might even better observe and interpret than the person herself, for her view is distorted by her self-conception and avowals. As a consequence, with respect to substantial self-knowledge, the person herself doesn't have a privileged position (epistemic or agential) vis-à-vis her attitudes, and will, presumably, be largely ignorant of them. As Schwitzgebel writes:

> If my attitudes – my beliefs and my values, especially – are not so much what I sincerely avow when the question is put to me explicitly but

[11] I focus on paradigmatic cases of care and withhold from discussions about the alleged possibility that a person might, e.g., care about X just a moment in time or very short period.

> rather what is reflected in my overall patterns of action and reaction, in my implicit assumptions, my spontaneous inclinations, then although I may have pretty good knowledge of the simple and trivial, or the relatively narrow and concrete – what I think of April's weather – the attitudes that are most morally central to my life, the ones crucial to my self-image, I tend to know only poorly… (Schwitzgebel 2012, 193)

The thought here seems to be that the possibility and prevalence of a gap between what a person avows and her patterns of action and reaction imply that she can only really know herself if she observes and interprets her actions and reactions (cf. Cassam 2014; Lawlor 2009; Schwitzgebel 2012). The role left for avowal is not entirely clear: is it just one piece of evidence amongst the rest of one's behavior? Is it an obstacle in achieving self-knowledge because it reflects the distorting lens of one's self-conception? Or does it serve as a contrast to the rest of our behavior, which helps us and others to reveal our own ignorance? Whatever the role left for avowal, whether it distorts our view of ourselves or helps us reveal our own ignorance, it isn't a necessary condition of self-knowledge. The objection to the view that avowal is necessary and has a unique status is thus that given that substantial attitude A is (also) reflected in one's patterns of action and reaction and given that these patterns might be, and often are, contrary to one's avowals, then one shouldn't rely on one's avowal of A in order to know whether one has A.

As far as I can tell, no one explicitly addresses this objection. There are two kinds of responses available in the literature, but both seek to safe the theory despite the objection, rather than address the objection itself. Matthew Boyle (2015), for instance, emphasizes that even if Moran's account doesn't apply to substantial self-knowledge, it still addresses *fundamental* questions about self-knowledge.[12] According to Boyle, Moran's project contributes to understanding the relation between having a mental life and being a subject with a first-person

[12] Other responses that accept that Moran applies only to trivial self-knowledge can be found in Schwenkler (2018), who claims that Moran addresses a paradigmatic form of self-knowledge, and Gertler (2016), who argues that trivial self-knowledge is the kind of self-knowledge that is epistemically distinct, which merits the philosophical attention given to it. What is quite striking in this respect is that such acceptance is, as far as I can tell, absent in Moran's view. He doesn't talk of trivial or substantial self-knowledge, but he does claim that his account applies not only to beliefs but also to emotions and intentions (Moran 2001, 64-5; 2012, 214; 2004, 471). Moreover, the examples that Moran turns to are often examples of substantial self-knowledge: for instance, the case of the analysand (2001, 93-5); akratic gambler (2001, 78-82; 162-3); the rakehell (2001, 174-187); and Fred Vincy (2001, 188-192).

perspective.[13] The other kind of response is to say that avowal is necessary but should be supplemented with other necessary conditions. In this vein, McGeer (1996; 2007) argues that a person's capacity to avow mental attitudes should be supplemented with a capacity for self-regulation. That is, a person's authority to avow a mental attitude depends on her willingness and capacity to regulate her performance accordingly, i.e., 'to bring [her] words and deeds into comprehensible alignment' (McGeer 2007, 87).

Where Boyle seeks to defend the importance of avowal despite the lack of application to substantial self-knowledge, and where McGeer seeks to undermine the starting-point of the question, namely that our avowals *will* resonate in our behavior if we regulate ourselves properly, I am interested in the importance of avowal *even in face of* the possibility of a lack of alignment between our words and deeds. What if we accept that (a) substantial attitudes are (also) reflected in one's patterns of action and reaction and accept that (b) these patterns might be, and often are, contrary to one's avowals, but *reject* that this means that (c) one shouldn't rely on one's avowals?

In what follows, I will argue that, despite (a) and (b), avowal remains necessary and still has a unique status in achieving self-knowledge of one's substantial mental attitudes. There are two ways to reject the conclusion that we shouldn't rely on avowals. First, I will inquire whether it is possible to deny *any* status to avowal in achieving substantial self-knowledge (section 4). For this seems to be assumed by those putting forth the objection: that a person's substantial attitudes can be "discovered" in her patterns of action and reaction – without any necessary role for avowal. Let's call this the *discovery assumption*. Secondly, the objection uses the possible gap between avowal and engagement as a reason to dismiss avowal's unique status in achieving self-knowledge. I will question whether this is a good reason to do so (section 5). This requires developing a view of the nature of the unique status of avowal in achieving substantial self-knowledge.

[13] As Boyle (2015, 346) writes: 'The idea is that, to understand the mind, we must understand subjectivity, and subjectivity is expressed primarily in a special mode of awareness of certain states: awareness of them from a standpoint one has precisely in virtue of being in those states.' Boyle is explaining his sympathy for a claim made by Sidney Shoemaker: '…it is essential to a philosophical understanding of the mental that we appreciate that there is a first-person perspective on it, a distinctive way mental states present themselves to the subjects whose states they are, and that an essential part of the philosophical task is to give an account of mind which makes intelligible the perspective mental subjects have on their own mental lives' (Shoemaker 1996, 157; quoted in Boyle 2015, 345).

4. Self-knowledge without avowal?

In this section, I will question the discovery assumption: the assumption that substantial mental attitudes can be discovered in patterns of action and reaction independently of a person's avowals. An influential example used to argue for the idea that one's prior engagement is evidence for having an attitude is Lawlor's example of Katherine, who achieves self-knowledge by *inferring* whether she desires to have another child through *internal promptings*, such as that 'she finds herself lingering over the memory of how a newborn feels in one's arms' and that '[s]he notes an emotion that could be envy when an acquaintance reveals her pregnancy' (2009, 47). Internal promptings are imaginings, fantasies, memories, emotions and sensations and, according to Lawlor, 'self-knowledge of desire is in routine cases a matter of self-interpretation of one's imaginings, where that self-interpretation is a causal inference to the best explanation' (2009, 62). But can internal promptings serve as evidence in the way suggested by Lawlor? Does Katherine really need to (provisionally) *discover a fact about herself*, namely whether she does or doesn't want another child (cf. Lawlor 2009, 57)?

Lawlor's case, as it is described, rejects the need for making an avowal: self-knowledge is acquired by paying close attention to one's internal promptings and then inferring which mental attitude best explains these inner promptings. Responding to the case of Katherine, Boyle makes clear that he doesn't take this to be a genuine possibility:

> A person can certainly realize that she wants another child by paying attention to her own thoughts and feelings in the way Lawlor describes, but is it really plausible to represent this as a matter of detecting some standing fact of the matter? Her feelings when she boxes up outgrown clothes and receives news of her friend's pregnancy are certainly indications of an incipient desire, but 'incipient' is important here. It is natural to imagine her also thinking of ways in which having another child would make it difficult to pursue other things she cares about. What she wants to know, presumably, is whether the decision to have another child is one she can genuinely embrace, and though 'inner promptings' may serve as indications of such a readiness, this is not simply a question of discovering what is already so but of reaching a settled attitude on the matter. To investigate this as if it were a matter for discovery on the basis of

> evidence sounds, even here, like alienation, or indeed like bad faith. (Boyle 2015, 344)

Now, Boyle claims, first, that acquiring self-knowledge of the desire to have another child is not a matter of detecting a standing fact, but of reaching a settled attitude on the matter. Secondly, he claims that the reason for this is that such self-knowledge amounts to knowing whether the decision to have another child is one a person can genuinely embrace. Although I agree with the first claim, I think the second misrepresents the case. Boyle seems to offer a revision of the original case: on his portrayal of the case, Katherine needs to *decide* to have another child instead of merely determine whether she has a *desire* to have another child. Having a desire to have another child is not the same as viewing the desire as something to pursue *all things considered*. Katherine might have the desire to have another child, even if she thinks having another child would conflict with her desire to pursue her career and therefore, she cannot fully embrace the *decision* to have another child. After all, not all things that one deems to be good can be pursued at the same time. Referring to decision (and its commitments) thus doesn't solve the question raised by Lawlor's example. Namely: why is it problematic to claim that internal promptings are *evidence* for having the desire? Can't we imagine that Katherine would experience internal promptings to such a degree that she cannot but infer that she has the desire?

Notably, Moran extensively discusses cases where doubt about the strength of one's decision (or commitment) undermines the decision (or commitment).[14] Such doubt sometimes seems to be required if one is to be 'realistic' about oneself, i.e., about one's character and capacities. In view of this, the question Moran (2001, 81) asks himself is how taking responsibility can be compatible with being psychologically realistic about oneself. What is the relation between an *avowal* and the *psychological facts of the matter*, i.e., the facts manifested in patterns of action and reaction? And here Moran argues that avowal is necessary for first-personal self-knowledge, because the psychological facts of the matter cannot form a *sufficient* basis for self-knowledge. As Moran writes:

> The assertion from the Deliberative stance that "I am not bound by my empirical history" is not in any way a denial that the facts of my history are what they are. It does not deny either the truth of these claims or

[14] See, for instance, as already mentioned in fn. 12, Moran's discussion of the akratic gambler (2001, 78-82; 162-3); the rakehell (2001, 174-187); and Fred Vincy (2001, 188-192).

> their relevance to the question at hand; but it does deny their completeness and, in a word, their decisiveness.[15] (Moran 2001, 163)

What does Moran mean by saying that the facts of the matter aren't *complete* or *decisive*? Understanding this is the same as understanding why a substantial attitude isn't simply revealed in one's patterns of action and reaction and will tell us whether to reject the discovery assumption.

Returning to the case of Katherine, the question to be asked is how she knows what her internal promptings, e.g., her envy, indicate. How does Katherine know that her envy reveals a deeper truth about herself and isn't due to, for example, having a grumpy day? The envy itself doesn't wear it on its sleeves whether it is a symptom of a deeper desire. After all, emotional episodes have different kinds of significance for a person. Among other things, a person may discard them for making a fuss about something insignificant, or she may experience them as an *expression* of what she cares about. Katherine's envy thus cannot be seen as *plain evidence* for her deeper desire.

Furthermore, in the description of the case, Lawlor writes that Katherine 'notes an emotion *that could be envy* when an acquaintance reveals her pregnancy' (2009, 47. My italics). We see another problem reflected in this passage, because before Katherine can take her envy to be indicative of her desire to have another child, she first needs to recognize her emotional reaction as envy. How does Katherine know that the emotion she feels is *envy*?

The problem taking shape here is that, like desire itself, internal promptings often are *mental attitudes* one needs to know about. Is it possible to argue that Katherine knows it is envy she feels because of yet other internal promptings that could be taken as evidence for it? Such a response isn't possible *ad infinitum*. As Moran writes, it is impossible to treat one's entire mental life as mere data: 'a person cannot treat his mental goings-on as just so much data or evidence about his state of mind all the way down' (Moran 2001, 150). At a certain point a person must recognize mental data as expressing her stance on the matter. The *symptomatic* value of mental data or internal promptings, i.e., their evidential status, depends on whether the subject takes them to be expressive of her perspective. For instance, if a desire to scream whilst being at the opera would pop up, I would immediately

[15] See fn. 8. See also, for instance, *Anscombe's principle* as formulated by Setiya (2011, 174): 'If A has the capacity to act for reasons, she has the capacity to know what she is doing without observation or inference – in that her knowledge does not rest on *sufficient prior evidence*' (my italics). See Falvey (2000) for an insightful discussion of the implications of such a claim in the case of intention and action.

disavow it and wouldn't see it as symptomatic of any of my deeper values.[16] As stated by Moran, '[a]t some point, I must cease attempting to infer from some occurrence to my belief; and instead *stake* myself, and relate to my mental life not as something of symptomatic value, but as my current commitment to how things are out there' (2001, 150).[17] That is to say that even if, in principle, it is possible to treat any mental attitude as a mere datum, it is impossible to treat my entire mental life as mere data: in each case of treating a mental attitude as datum, I must also, at some point, *stake* myself. Hence, Katherine's internal promptings are evidence for having the desire to have another child only if they are related in the relevant way to 'internal promptings' *that she avows*.[18]

The upshot of this is that it is misguided to portray the subject's relation to her 'internal promptings' (or her *engagement* with the object of her care) as merely passive. Internal promptings aren't just *given facts* about the subject that she can discover, experience or note, but are, at least some of them, an expression of the subject's commitments to the world at large. A subject's engagement with the object of her care is not something that just miraculously happens; rather it is, at least in part, an expression of her agency. After all, the engagement itself also consists of attitudes that involve commitments to a certain grasp of the world. As I have hoped to show, this means that a subject's prior actions and reactions cannot be taken as plain evidence, sufficient to determine whether she has a specific mental attitude. Avowing a mental attitude cannot float free from one's patterns of action and reaction, but that doesn't mean that these patterns can *replace* avowals completely.

5. The significance of the gap

To know that a person cannot achieve self-knowledge of her substantial mental attitudes without avowal is not yet to understand the precise function of avowal in

[16] I owe this vivid example to Fleur Jongepier.

[17] Cf. Moran (2001, 121-4). If I would treat all my mental life as mere data, I could not even arrive at a *conclusion* about my mental attitudes: 'The radical abrogation of first-person authority means that he cannot take for granted that the conclusion he arrives at just is, now, what he genuinely believes about the matter. Thus, his problem is not only that the current of his true beliefs and feelings runs somewhere out of sight of his consciousness, but also that this current seems to run its own course and have nothing to do with his explicit thinking about the people and things his feelings are supposedly directed upon' (2001, 123).

[18] Wouldn't Lawlor simply reject that Katherine needs to *avow* her internal promptings and instead claim that they might just *feel* as expressive of her perspective? Can't she just turn to phenomenology? In response, I take it that an internal prompting can only feel as expressive of one's perspective if it feels related to one's grasp on the world, i.e., if it is related to one's commitments. This is precisely what avowal is about.

achieving such self-knowledge. What is the positive role of avowal? And how can this role be reconciled with the possibility and prevalence of a gap between a person's avowals and her patterns of action and reaction?

It seems abundantly clear that it is possible and prevalent that we fail to live up to our avowals. I say I believe that taking care of the environment is important but fail to actually take care of the environment (or at least choose the alternatives that do less damage to the environment). I say I care about my health but fail to establish healthy habits. I say I care about my job but find myself struggling to see the value of what I do. In line with these examples, it seems that the avowal of what I care about cannot be taken to be an expression of a commitment that I *already have*, but rather is something more *provisional*, like a pledge to try to be committed in that way. Whether I actually have the commitment then seems to be a question that can only be answered either by checking with my future engagement itself or by knowing how likely it will be that I will be engaged in the appropriate way (e.g., I have done so in the past; I have good self-regulation skills). Let's call this the *skeptical picture*, where one's avowal of one's care is to be corroborated in (the likelihood of) the relevant kind of future engagement if it is to amount to knowing one's care. Such corroboration is needed because of the possibility and prevalence of a gap between avowal and engagement.[19]

Although I see the pull of the skeptical picture, I think it is mistaken. Even if we accept the possibility and prevalence of a gap between a person's avowals and her patterns of action and reaction, it isn't thereby determined what such a gap implies. It seems often assumed that such a gap implies ignorance or unreliability, but this isn't a necessary consequence. I will argue that mistakes might actually be inherent in the kind of process or the kind of capacity at issue.[20] I will discuss three reasons against the skeptical picture. These reasons concern 1) the provisional nature of avowals, 2) the significance of a gap between avowal and engagement, and 3) the kind of engagement involved. These different considerations might feel disconnected, but together form the basis of a different picture of the role of avowal in which it has a unique status.

The first reason against the skeptical picture concerns the provisional nature of avowals. That avowal has a provisional quality is, in my view, a direct

[19] The "skeptical picture" owes its label to the argument from illusion, which assumes, roughly, that the possibility of being wrong (being under the sway of an illusion) implies the need for external justification. Cf. Dancy (1995).

[20] The idea that exercising capacities, especially "normative" capacities, involves the possibility of failure is not a new idea. See, for instance, Korsgaard (1996). However, the ideas presented about the significance of this possibility and of making mistakes in light of achieving substantial self-knowledge are, as far as I can tell, novel.

consequence of the relation a person has to her own mental life. For recall that this relation is one where she sees her intentional mental attitudes not as given facts but as expressive of her grasp of the world – a grasp that is always evolving as one's circumstances change. Avowing one's mental attitude isn't supposed to alter this relation; rather, it merely expresses it – or we might say, pays "tribute" to it. A mental attitude is not, after a person avows it, suddenly considered as a given fact. Rather, a person will keep seeing the attitude as expressive of her commitment, and consequently as something she needs to sustain. For instance, if one makes a resolution to write a grant proposal, one not only needs to take action to actually write the grant proposal, but the resolution itself needs to be sustained throughout the time one needs to finish the grant proposal. Similarly, caring about one's job is not all of a sudden turned into a given psychological fact about a person if she avows it. By contrast, it remains something that she needs to sustain. Hence, given the idea that mental attitudes are expressive of one's commitments, avowal *should be* provisional.

This also has consequences for the relation between avowal and other agential capacities, such as self-regulation. It seems as if self-regulation doesn't only supplement avowal, as for instance in McGeer's (2007) proposal, but can also undermine one's capacity to avow one's mental attitudes. According to McGeer, self-regulation involves taking an instrumental stance vis-à-vis one's own mental attitudes: one sets oneself to the task to make an avowed mental attitude fully one's own (cf. McGeer 2007, 90). To be able to envisage such a task, one must think about the mental attitude as something settled – only if there is a settled endpoint in view, can one take up the means to get there. Hence, taking up such an instrumental stance involves taking one's avowed mental attitude as given.

In the case of writing the grant proposal, self-regulation is needed to arrange the external and internal circumstances in such a way that one can fulfill one's resolution, e.g., create an environment in which one is able to concentrate, make sure one feels supported by one's colleagues to try out new ideas, let oneself not be overcome by anxiety because the competition is so high, etcetera. But insofar as this implies viewing one's mental attitudes as given, self-regulation is not the kind of activity through which one stays committed to one's resolution.

What's more, it might sometimes even hinder one to notice whether one still is to be committed to one's resolution. If one is too focused on finishing that grant proposal, one might miss that writing the grant proposal is no longer valuable to oneself, i.e., one might miss how one's grasp of the world at large evolves. As McGeer herself writes in the penultimate paragraph of her article, if one is too focused on self-regulation in order to stick to some resolution or commitment, one may regard

one's own reactions, i.e., reactions that may reveal something about one's grasp of the world, as "wayward tendencies" that one should overcome, instead of taking them as expressive of one's stance. McGeer refers here to 'some of the American soldiers who, against their own feelings of anxiety or revulsion, ended up torturing prisoners in Iraq's Abu Ghraib' (2007, 104). One problem with the soldiers, as McGeer observes, is that they *regulated* these negative feelings, whereas they should have taken them at face value: they expressed their grasp of the horrific nature of what happened in Abu Ghraib.[21] The instrumental stance belonging to self-regulation might thus hinder one to see one's mental attitudes as expressing one's view on things and thereby hinder one's capacity to avow one's mental attitudes.

The second reason speaking against the skeptical picture has to do with the significance of a possible or real gap between avowal and (future) engagement. The problem is that observing a person's engagement isn't decisive in determining 1) whether the gap between avowal of one's care and future engagement actually constitutes a *failure* to live up to one's avowed care, nor 2) whether that failure constitutes a *failure to know* one's care. First, it cannot be concluded that a person fails to live up to her avowed care by observing her engagement because a gap between avowal and engagement might also be the result of a change of mind (or, for "care" perhaps better to say: a change of heart). For instance, I could believe that *p* and know this of myself, and after a change of mind (due to, for instance, learning new information about the issue) believe that *q* instead of *p* and still know this of myself. Similarly, I could care about X and, after a sufficient amount of time, have a change of mind (or heart) and stop caring about X. This need not impugn that I know I cared about X.[22] Whether a gap between avowal and (future) engagement is a failure or a result of a change of mind is not something that can be observed in the engagement itself. This reflects the situation of knowledge of one's actions: if I get up to make green tea and, while I am making tea, make lapsang souchong instead,

[21] I don't mean to claim here that what McGeer describes as Moran's deliberative ideal would be sufficient in the case of the American soldiers to have retained from torture. I think that the social influence (and of course the influence of being at war) on self-knowledge and on one's ability to recognize one's grasp of the world is substantial. But I see it more as *part of the struggle* to avow one's attitudes (cf. Pippin 2005). More on this struggle further on in this section.

[22] This is not to say that any change of mind (or heart) is acceptable. For instance, 'sufficient amount of time' is important here, especially for substantial self-knowledge. If the period is too short, it becomes doubtful whether my care for X is *genuine*. Or if the change of mind happens randomly, i.e., for no reason at all, this might also raise the suspicion that I didn't genuinely care for X. Additionally, it would be dubious whether I actually cared for X if I would change my mind (or heart) successively or very often. In situations like these, having a change of mind (or heart) does seem to constitute a failure to live up to one's avowal.

you have to ask me whether I made a mistake or whether I changed my mind (cf. Falvey 2000).

Secondly, it cannot be concluded from failing to live up to one's avowed care that one actually fails to know one's care. There must be a difference between failing to live up to and failing to know one's care, because failing to live up to a care (and the commitments inherent in caring) presupposes having it. Again, I don't think any failure (or any amount of failure) is permissible if one is to count as having the care. For instance, failing on too many or too important occasions undermines caring about something. But like the difference between a change of mind and a failure, this difference isn't observable in the engagement itself. There isn't a clear division between a failure to live up to one's care about X and a failure to care about X (and thus a failure to know it). Rather, what counts as a certain kind of failure is, in my view, the result of a complicated process in which both the person herself reconsiders whether she cares about X (and thus whether X is important for her) and negotiates with others whether her engagement with X, including failures to have the right kind of engagement, suffices for caring about X.[23]

Following these points, I find myself in disagreement with Schwitzgebel (2012), who suggests that failing to live up to an avowal should be understood as a failure to know oneself. He gives the following example:

> I say I value family over work. When I stop to consider it, it seems to me vastly more important to be a good father than to craft a few more essays like this one. Yet I'm off to work early, I come home late. I take family vacations and my mind is wandering in the philosopher's ether. I'm more elated by my rising prestige than by my son's successes in school. My wife rightly scolds me: Do I really believe that family is more important?

For Schwitzgebel, this example demonstrates our self-ignorance: we say (or avow) one thing but behave to the contrary. I don't subscribe to this implication and think it leaves several assumptions unspecified. Even if we agree with Schwitzgebel that the things he describes constitute failures to live up to valuing family over work, and thus agree that his wife *rightly* scolds him, do we then *know* whether Schwitzgebel

[23] This latter condition, i.e., the negotiation with others, is part of the social dimension of self-knowledge, which, as already admitted, I cannot do full justice in this paper due to lack of space and complexity. See McGeer (2007), especially for the social dimension of developing the requisite capacities for first-person authority. And see Pippin (2005, 309, 318-322) for a discussion of the influence of "negotiation with others."

actually values family over work or vice versa? Drawing such a conclusion would assume, first of all, that these things can be taken as *plain evidence* for his values. As argued in the previous section, such evidence isn't sufficient to determine what Schwitzgebel values more. Secondly, such a conclusion assumes that his mistakes are unambiguous signs of his self-ignorance. As I just argued, however, his failures could also indicate, not self-ignorance, but that he just fails to live up to his values. Which of the two it will be is a matter to discuss, not for us, philosophers, but for Schwitzgebel and his wife: he needs to reconsider and negotiate with his wife whether he values family over work and, importantly, what kind of engagement with his family is actually demanded by the care he has for his family. For instance, is it really problematic that, even during family vacations, his mind is wandering in the philosopher's ether? Or that he spends time on writings essays? Answers to this kind of questions are, I think, not clear-cut, but a matter of negotiation both with oneself and with others. This brings me to the next point against the skeptical picture.

The third reason counting against the skeptical picture concerns the kind of engagement that is involved in caring about X and the way in which one knows about this engagement. In the skeptical picture, it is possible to check whether a person is engaged in the right way (in the future), which presupposes that there is some set standard how she should be engaged. In opposition, I want to argue that understanding the right kind of engagement belonging to caring about X, especially what the right kind of engagement is *for a particular person*, cannot be the result of theorizing about what it means to care about X, but can only come about by trying to care about X. Knowing what it means for a person to care about X, and thus what kind of engagement is involved, is something only she, and only by trying to care about X, can understand.

To see what I'm trying to get at, consider Robert Pippin's analysis of what is involved in knowing one's practical identity. In an essay on Proust and self-knowledge, Pippin describes Marcel's (search for) knowledge of being a writer in the following way:

> The young Marcel considers himself, from very early on, a writer; that is his self-understanding; and he is very much trying to become who he believes he is, trying to become a writer. And this is indeed portrayed as a struggle… For a very long time… Marcel is a writer who does not write or writes very little as he struggles to understand how a writer lives, how one responds to and tries to understand the people around him "as a writer would" and struggles to find out whether he

> can ever become in reality, however much he actually writes, "a real writer." (Pippin 2005, 315)

Pippin portrays Marcel's struggle to become who he considers himself to be – to make his self-conception true – not only as a struggle to actually write but also as a struggle to understand what it means to be a writer. Especially, it is a struggle to understand, amidst Marcel's own and society's expectations of how a writer should live, what it means for *himself* to be a writer. Such an understanding, as argued by Pippin, cannot be achieved by mere theoretical means – not by contemplating the life of a writer nor by searching for one's own "writerly essence" (2005, 331). It is instead a matter of trying to be a writer: in the act of writing, in failing to write, and in negotiating with others what it is that one is doing, one can start to understand what it means for oneself to be (or failing to be) a writer.

In my view, the struggle to care about X should be understood in a similar vein. There are two ways in which the meaning of caring about X depends on trying to care (or in trying to live up to one's care): only in trying to care will the commitments inherent in caring about X become apparent, on the one hand, and be truly understood, on the other. Let me explicate these claims with an example.

Suppose Joan avows that she wants to spend the rest of her life with David[24] (for brevity's sake, let's say she wants to be *married* to David). And suppose we say that she should assess the likelihood of her staying faithful to David and being able to spend her whole life with him, in good and bad times, so as to know whether she *really* wants to be married to him. On what would she base such an assessment? Her commitment to David in the past? Her resilience in dealing with temptations and setbacks? Obviously, as portrayed in the previous section, these sorts of things matter. Joan's avowal that she wants to be married to David cannot float free from the patterns of her actions and reactions. But do these sorts of things provide information about what it means to spend the rest of her life with David? What it means to grow old together? What it means to take care of each other for the rest of one's life? Joan will have ideas about these things, she will have imagined her future with David repetitively, and she will harbor expectations about how her life together with David will be. In other words, her wanting to be married to David is tied to a conception of how her life with David will be, and with a self-conception of the person she will be as having David as her husband and as being his wife.

[24] David could, of course, also be another woman. Being of different gender, however, makes it easier in writing, because one can refer to "he", "she", "him" and "her".

But in trying to be that person and in trying to have the life she imagined, she will unquestionably experience tension between her expectations and how it turns out to be. She might doubt whether she can endure the fights they have and, next, doubt whether her doubt can be part of wanting to be married to David. She might experience difficulties in accepting David's growing fondness of taking long solitary walks. This faces her with the questions what amount of experienced difficulties in marriage will be acceptable *for her* and when she will stop wanting to be married to David. Such questions, however, aren't answered by theoretical reflection but by actually experiencing the mentioned doubt and by returning to one's sense of commitment (or failing to), or by actually experiencing the difficulties and finding a way (or failing to find one) to accept or deal with them. Such experiences and challenges put into question whether Joan sees herself as the kind of person who is committed to being married to David under the current circumstances. Hence, in both these ways – i.e., in understanding which commitments are inherent in wanting to be married to David and in truly understanding the commitment itself – grasping the meaning of wanting to be married to David can only come about by trying to live the rest of her life with him.

Knowing about the engagement involved in caring about X thus requires trying to care about X. But what is the role of avowal in all this? The role of avowal (either verbally, in inner speech or in thought) is tied to having a self-conception, and in making oneself sensitive to questions about the right kind of engagement. I first need to avow my care and have a particular self-conception of what I care about before I can either sustain or fail to sustain it. Only through avowing my care about X, can my commitments to X be challenged, whereby I come to experience what caring about X asks of me and whether I can be *and want to be* the person meeting those demands. If Joan doesn't avow that she wants to be married to David (or has the self-conception that she does), she might actually be spending her life with him and 'show' the behavior that is deemed appropriate, but the relation to her relationship with David is not the same. She cannot be challenged or take responsibility in the same way. For instance, if she starts to have feelings for someone else or if she accepts a job-offer abroad, and as a consequence her relationship with David changes, there is no immediate reason to criticize Joan. Without avowal, she doesn't take responsibility for staying committed to David, nor for developing an understanding of what it means to spend the rest of one's life with him.

Let's take stock. The skeptical picture, where one's future engagement plays the role of evidence in knowing whether one cares about X, is mistaken in my view, because avowal actually should be provisional, because one cannot observe

whether a gap between an avowed care and engagement is due to a failure to have the care, and because the required engagement regarding X isn't set in stone, but rather is something the person herself must come to understand in trying to care about X.

The alternative picture taking shape here is that the possibility and prevalence of failing to live up to one's avowals – the gap between avowal and engagement – is not necessarily a sign of ignorance (and so a sign that one's avowal should be corroborated in one's behavior), but part and parcel of the *kind of thing that avowing one's care is*. The possibility of failure is part of avowing one's care precisely because caring involves expressing one's commitment through avowal. And similar to acting, planning, intending, etc., an agent can fail in being committed: an agent can act badly, fail to fulfill her plans, fail to act on her intentions and, analogously, fail to fulfill her commitments. Under certain conditions, e.g., if one fails on too many or too important occasions, such failure might indicate a lack of care and thus be a sign of self-ignorance. However, this cannot be decided by some independent criteria, but is something the person herself must negotiate, both with herself as well as with others.

This alternative picture brings us back to the agential aspects of self-knowledge. Agents do not merely suffer what happens but lead their life. This essentially involves acting and reacting in light of some conception of who one is committed to be. Acting in light of a particular self-conception implies the possibility that the action fails to be in line with the self-conception. Therefore, being an agent also means being able to *fail to lead one's life*, and indeed, sometimes, just suffer what happens.[25] In this picture, the role for avowal is first and foremost to have such a self-conception. To avow is to take up, what we might say, *the burden of agency*: to take a stance, commit oneself and have a self-conception; to thereby put oneself at risk of making mistakes; and through both these things, to seek deeper understanding of who one is.[26] This brings us back to the main question of the paper:

---

[25] Here, I am inspired by and paraphrasing Pippin (2005, 309): 'Being the subject of one's life, a subject who can lead a life rather than merely suffer what happens, who can recognize her own agency, the exercise of her subjectivity, in the deeds she produces, also means *being able to fail to be one*.' However, there is an essential difference between the two ways of phrasing: where Pippin sees the concept of agency as something one can fail to be, I think that the concept of agency is connected to failure by exercising one's agency, i.e., through acting, intending, committing, etc. One cannot *fail to be an agent*, but, *as an agent,* one can *fail to lead one's life*. I don't know who would do the failing otherwise.

[26] Compare again Pippin (2005, 317): 'Put one final way, the problem I have called Marcel's "becoming who he is" amounts to his becoming a determinate agent, someone who leads his life, both carries the past into the future in a certain way and does so, acts, in light of some conception of the subject he is struggling to become. But this is mostly manifested by a kind of *via negativa*, the often palpable sense in the novel of the great and almost intolerable burden of the demands of such agency and the sweet pleasures to be gained by avoiding such a burden.'

does avowal have a necessary and unique status in achieving substantial self-knowledge? What I have hoped to show in my discussion is that avowal is, *par excellence*, essential to substantial self-knowledge.

6. Concluding remarks

In this paper, I have taken issue with an objection against the unique status of avowal. The objection is that, given that (a) substantial attitudes are (also) reflected in one's patterns of action and reaction and that (b) these patterns might be, and often are, contrary to one's avowals, (c) we shouldn't rely on our avowals. I have argued that, despite (a) and (b), avowal remains necessary and still has a unique status in achieving self-knowledge of one's substantial mental attitudes.

I have focused on the attitude of care and considered the relation between avowing one's care about X and being engaged with X in the right way, i.e., having the right kind of pattern of actions and reactions. These patterns aren't decisive in determining whether a person cares about X, because they only have significance as evidence if, at some point, she does commit herself to a certain grasp of the world. Nor can a person's (estimated) future engagement be used to corroborate whether she actually cares about X, because, in the end, only she can determine (in negotiation with others) whether 1) she is living up to her care, 2) whether a gap between her avowal and engagement constitutes a failure, and whether that failure is a failure to live up to her care about X or a failure to actually care about X, and 3) what kind of engagement is actually required in caring about X.

On the alternative picture, the possibility of failing to live up to an avowal is part of what an avowal is, especially in the case of substantial self-knowledge. For knowing one's substantial mental attitudes is best viewed as a struggle: in the case of care, a struggle to try to care about X by avowing the care and act on that avowal – that is, to act in light of a self-conception – thereby putting oneself at risk of being challenged and of making mistakes. Engrossing oneself in this struggle is essential to achieve self-knowledge of caring about X because it is necessary for understanding what it means for oneself to care about X.

CONCLUDING REFLECTIONS

The aim of this dissertation has been to contribute to an understanding of self-knowledge within a moral psychological framework, where the connections to personhood, moral psychology, and (mental) agency are recognized as crucial to understanding the nature of self-knowledge. I have given critical analyses of Moran's account of self-knowledge, of the different responses to the two topics problem, of the nature of reasoning and of the importance of a person's active relation to her own substantial mental attitudes. Taken together, they help to address the underlying question of this dissertation: what is transparent self-knowledge and why should we adhere to it?

In these concluding reflections, I will revisit the main themes put forth in the introduction. I will discuss the limits of transparency accounts in section 1 (related to the themes "Transparency", "Two Topics", and "Attitudes and Scope"). In section 2, I want to reflect on the role that the *form* of transparent self-knowledge plays (related to the themes "Agency" and "Approach"). Finally, in section 3, I will consider how these themes are related to scientism and the scientific perspective on self-knowledge.

1. The limits of transparency

The overarching result of my inquiry is that a person's active relation to her own mental attitudes, as introduced by Moran and as further developed in this dissertation, is central to transparent self-knowledge. That is to say, to the kind of self-knowledge that respects a certain kind of transparency, namely, the connection between having a mental attitude and viewing the world a certain way. Whether a mental attitude is trivial or substantial, the fact that our attitudes express our own stance implies that the question of taking responsibility has become pertinent to all of them. When self-ascribing an attitude, we either take or *fail to take* responsibility for our attitudes by avowing (or disavowing) them. And because of this, achieving the kind of transparent self-knowledge at issue involves manifesting one's agency:

for only if a person takes the responsibility to avow her mental attitude will she achieve transparent self-knowledge. The resulting picture of the kind of self-knowledge at issue, i.e. transparent self-knowledge, is thus that it is agential, first-personal, and transparent, at least if we take the latter to mean that it respects the connection between having a mental attitude and seeing the world in the relevant way.

Characterizing transparent self-knowledge in this way is distinct from identifying any *transparency procedure*. Whether achieving transparent self-knowledge should be explicated as going through a transparency procedure is a different matter. Even if self-knowledge is often thought to be transparent by virtue of such a procedure, it seems exceedingly difficult to explain the picture of self-knowledge just sketched in terms of a transparency procedure.

Formulating a transparency procedure is to identify the way in which a person achieves self-knowledge. Such a procedure can be called transparent if it follows the sense of Evans' notion of transparency (as distinguished from Cartesian transparency and diaphanousness). That is to say, if the procedure involves answering an inward-directed question about one's mental attitude by answering the relevant outward-directed question about the content of the attitude. Or put differently, if the procedure involves the idea that the subject learns of her mental attitudes by attending to their objects. I have presented two arguments against thinking that a transparency procedure depicts how we come to acquire transparent self-knowledge.

The first argument concerns the scope of a transparency procedure. In this dissertation, I have looked at the scope of Moran's transparency claim. No matter how the claim is distilled from his work, we seem to get into trouble regarding its scope. Should Moran's transparency claim entail "considering reasons," whether these are reasons relevant to answering the outward-directed question or reasons justifying one's answer to that question (presented as requirement two and three in Chapter 1), its scope is fairly limited. A person has and knows she has many beliefs for which reasons are not immediately available, such as basic or anti-skeptical beliefs, or for which it would be overly laborious to have to go over one's reasons again before being able to self-attribute them, such as long-standing beliefs. Thus, the scope of Moran's transparency claim diverges from what we think the scope of a transparency claim should be – it cannot account for all classes of belief of which we, intuitively, do have self-knowledge.

Should Moran's transparency claim not entail referencing reasons, but only involve answering the inward-directed question *by answering* the outward-directed question (presented as requirement one in Chapter 1), its scope is too wide rather

than too narrow. For, on this reading, the claim seems to apply to almost all mental attitudes. Included are attitudes, such as obsessive beliefs, that seem to be paradigm cases of alienated attitudes. Therefore, this minimalist reading of the claim doesn't seem to meet the scope that Moran himself has set for transparency, namely that it shouldn't apply to alienated attitudes.

Moreover, it appears to be very difficult to delineate an outward-directed question for attitudes other than belief (see Chapter 4). The result of my reflections on the case of emotion is that whereas the question whether to believe that *p* seems to be fully transparent to one single question, namely the question whether *p* is true, there isn't such a single question in the case of (at least some) other attitudes. While I don't want to claim that this implies that we cannot have transparent self-knowledge of these other attitudes, I do think it places strong doubt on the feasibility of Moran's transparency claim (the *procedure* described in these claims) in the case of other attitudes. The claim doesn't adequately describe how a person arrives at transparent self-knowledge of attitudes other than belief.

According to the first argument then, either the scope of Moran's transparency claim doesn't match Moran's delineation of what the scope is supposed to be or its scope is more limited than it should be. Of course, this argument, as it was advanced in the dissertation, holds only against *Moran's* transparency claim (although many parallel arguments exist against other transparency claims). The second argument, however, has been elaborated for a wider variety of views.

The second argument has to do with the Two Topics Problem (TTP). None of the discussed transparency procedures seem to be able to solve TTP (see Chapter 2). How can a person learn of her mental attitudes by attending to their objects, if those objects neither imply nor indicate anything about them? The truth of *p* just isn't evidence for nor necessitates one's belief that *p*. What could justify using *p* as an epistemic basis to self-ascribe the belief that *p*? What I have found in the case of believing that *p*, is that none of the discussed accounts is able to explain that the subject's commitment to the truth of *p* relates to or confers epistemic justification on her self-ascription of the belief that *p*. Attending to *p* just doesn't seem to do all the work that is needed to arrive at transparent self-knowledge: a self-attribution which is *transparent* because one is committed to the view purported by the self-attributed attitude, and which is *knowledge* because it is epistemically justified.

One of the accounts discussed was Byrne's inferential account of self-knowledge. Based on my argument on the nature of reasoning (in Chapter 3), I argued that his assumption that inference from a premise entails belief in that premise, which is required for the epistemic justification of the inference, is

unwarranted. What I have argued is that the orthodox view of reasoning (which I dubbed the attitude view) misconstrues reasoning as a mental process that involves a change in attitudes. Not all reasoning involves such a change in attitudes and hence, reasoning doesn't necessarily involve a change in attitudes.

I would like to add to these considerations that the argument on the nature of reasoning provides good reasons to think that transparent self-knowledge, but also self-knowledge in general, cannot be explained by inferential accounts of self-knowledge. If reasoning doesn't necessarily involve a change in attitudes, then we can make a distinction between reasoning with and reasoning without a change in attitudes. Let's call instances of the latter hypothetical reasoning and instances of the former, where one does believe the premises and conclusion of one's reasoning, categorical reasoning. I take it to be an uncontroversial datum that we have the ability to know when we reason hypothetically or categorically. If this is indeed the case, and if my argument that there is a distinction between hypothetical reasoning and categorical reasoning holds, then self-knowledge is prior to reasoning instead of the other way around. This would imply that inferential accounts of self-knowledge in general would be undermined.

The argument for this claim is the following. If reasoning is supposed to give us self-knowledge, as inferential accounts of self-knowledge claim it does, then the least we need is that through reasoning we know that we believe the conclusion of our reasoning. But if each instance of reasoning might either be a piece of reasoning involving or lacking belief in the premises and conclusion, then the fact that we are reasoning and arrive at a conclusion doesn't tell us anything about our beliefs regarding the premises and conclusion. Nor could we refer to another piece of reasoning to solve this lacuna, for there too, one's belief in the conclusion isn't implied in making the inference: if one is to believe the conclusion of *that* piece of reasoning, one would need yet *another* piece of reasoning to know of that belief, and infinite regress ensues. Hence, if we accept my argument that reasoning doesn't necessarily involve that one believes in the premises and conclusion of one's reasoning, then self-knowledge cannot be based on reasoning – at least not all of our self-knowledge. And as a corollary, this means that our ability to know that we believe the conclusion of our reasoning presupposes self-awareness of one's beliefs.[1]

Leaving these considerations on the difficulties of inferential accounts of self-knowledge aside, let me return to the question of the limits of transparency

[1] In addition, I think that something similar holds for deliberation. Making up one's mind by considering reasons presupposes awareness of what one takes to be a reason.

procedures. Because of the problem of scope and because of TTP, it seems that Evans' notion of transparency, including the transparency accounts based on it, cannot fully explain the source of transparent self-knowledge. One might think that saying that transparent self-knowledge isn't the result of going (only) through a transparency procedure implies a contradiction. But I want to suggest a different conclusion. We need to distinguish clearly between *transparent self-knowledge* and a *transparency procedure* to arrive at self-knowledge. Where the former expresses something about the nature of the self-knowledge at issue, the latter aims to identify an epistemically justified method of moving from thinking about the world at large to a self-ascription of a mental attitude. The way I have been explicating the nature of transparent self-knowledge, namely as self-knowledge that manifests the connection between mental attitude and view of the world, it isn't necessarily tied to a transparency procedure. Even if the way in which one achieves transparent self-knowledge is not fully transparent, there is no immediate implication that there isn't another way in which such self-knowledge might be achieved.

What could be an activity to achieve transparent self-knowledge that isn't a transparency procedure? As inherent to the idea of transparent self-knowledge, it cannot be the result of any form of self-observation. Additionally, the discussion of TTP has made clear that a person cannot achieve transparent self-knowledge by inference (Chapter 2). That same discussion has yet also made plausible that the kind of transparent self-knowledge at issue can only come about if we presuppose or postulate a form of *attitudinal awareness*.[2] Obviously, this brings with it a certain tension. Transparency, after all, centers around the idea of "gazing outward." How could transparent self-knowledge be compatible with a form of attitudinal awareness? Such compatibility would seem to depend on the way in which this form of awareness modifies a person's relation to her own mental life. It shouldn't impugn on, most importantly, the relation between attitude and commitment. Put differently, we might say that the compatibility of transparent self-knowledge and the involvement of attitudinal awareness requires the latter to meet certain transparency requirements. For instance, the requirement that any uncertainty in one's self-ascription should direct one's attention back to the content of one's self-ascribed state (i.e., gazing back at the world). In my view, future research on this topic should thus focus on which transparency requirements attitudinal awareness should meet if they are to be compatible.

[2] Chapter 2 showed that different kind of solutions to TTP postulate a form of attitudinal awareness. For example, phenomenal awareness (C. Peacocke), contrastive awareness (A. Peacocke), awareness that comes with a proximal intention (Roessler), awareness that springs from the structure of deliberation (Moran), and awareness of the mode of presentation (Boyle).

2. The form of transparent self-knowledge

If transparency accounts face all these difficulties, one might wonder why again we needed transparent self-knowledge in the first place. If we don't use a transparency procedure to arrive at self-knowledge, but possibly a kind of attitudinal awareness, why does it make sense to hold on to the concept of transparent self-knowledge? This question appears even more pertinent given the weight attached to the involvement of agency and avowal in the understanding of transparent self-knowledge as advanced in this dissertation. If the involvement of agency and avowal is not manifested in some kind of activity to arrive at self-knowledge, then how must we understand their contribution? Claiming that it makes sense to hold on to the concept of transparent self-knowledge depends, as I will try to explain, on understanding the *form* of transparent self-knowledge. My suggestion here is that the key to understanding transparent self-knowledge – and its indispensability for understanding our mental lives – isn't to be found in any transparency procedure but in its form. Such understanding seems to require an analytic Aristotelian approach.

Before explaining why, I think we need analytic Aristotelianism to understand transparent self-knowledge, let me first return to the basic motivation for thinking there must be such a thing as transparent self-knowledge. The motivation is that, in order to make a genuine distinction between the first-person and third-person perspective (and thereby between self-knowledge and other-knowledge), we need the concept of transparent self-knowledge. What sets the first-person stance apart is, first and foremost, the connection between having a mental attitude and grasping the world in a certain way. This is reflected in Evans' observations about transparency[3], but also, for instance, in the way we treat our own and each other's self-attributions. Following Moran, I take it that these "phenomena of transparency" require us to recognize the person's active relation to her own mental life. That is, from a first-person perspective, making a self-attribution is more than merely recognizing a psychological fact about oneself, for it requires one to take a stance. Even if one tries to evade taking responsibility for what one holds to be true, to be done, valuable, and worthwhile, the question what one's current commitments are stares one in the face. If we want to do justice to these "phenomena of transparency" and agential connection between mental attitudes and commitments, we need the concept of transparent self-knowledge.[4]

---

[3] And in Moore's paradox, which seems to expose a phenomenon that mirrors Evans' transparency.

[4] Other philosophers (and scientists), such as Carruthers and Dennett, would want to deny that any of the apparent differences between the first-person and third-person perspective amount to *substantial*

But the question remains how we should understand the distinctiveness and viability of transparent self-knowledge. This is where I think the analytic Aristotelian approach would be helpful. I have introduced analytic Aristotelianism in the discussion on the nature of reasoning (Chapter 3). Contrary to the attitude view of reasoning, it seems to be more fruitful to say that each instance of reasoning, in all their different varieties, involves a judgment of the specific form *p as following from q*. In our quest to analyze everything to its core, we might sometimes pull things apart that belong together. The problems we run into in analyzing reasoning as an activity where there is a certain relation between *different beliefs*, stimulate the view that reasoning is first and foremost *one thing*: namely recognizing the truth-connection between two statements (i.e., making a judgment that *p as following from q*). This is the *kind* of thing that reasoning is; its *logical form*. The logical form of a concept consists of the form of thought or the form of judgment that underlies the concept and it refers to what can be predicated of the thing in question, say X. The logical form (or *structure*) is revealed by analyzing the things that can be said or asked about X, and thus by analyzing our practices and abilities regarding X. I have argued that we need to understand reasoning in this way especially because of two reasons: first, it seems impossible to analyze reasoning in terms of sufficient and necessary conditions (or in terms of an explanatory essential feature or property), and secondly, because this seems to be the only way in which the extensive variety of instances of reasoning form a unity.

The suggestion I want to make is that something similar is true of transparent self-knowledge. Here too, each attempt at formulating necessary and sufficient conditions meets with numerous counterexamples. Here, too, separating the self-ascription from the commitments inherent in the self-ascribed attitude introduces a problem, namely TTP. And here, too, it seems that the wide variety of instances of transparent self-knowledge have something in common. What is this commonality? As evinced in the difficulty of finding an explanatory single shared feature or condition, the unity must consist in something else. Analytic Aristotelianism helps us to envisage what this something else might be.

This brings us back to Moran's remarks on transparent self-knowledge being a distinct *category*. As it was characterized from the start, transparent self-

differences between them. What would speak in favor of their position is that there appears no way to make sense or to give a unified understanding of the differences. In face of the "threats" to self-knowledge from a scientific, or perhaps "scientistic" perspective, it is therefore important to delineate the way in which the concept of transparent self-knowledge explains and unifies these different observations about the first-person perspective. In other words, if we can't make sense of transparent self-knowledge, then giving up on the idea of the distinctiveness of the first-person perspective becomes a more viable alternative.

knowledge is distinct, first of all, because it is knowledge of, say, one's belief that *p* *in which* one takes it to be the case that *p*. 'Genuinely transparent self-knowledge,' as Boyle (2019, 15) puts it, 'is not merely arrived at *by* considering whether *p*; it remains a mode of knowing *in* which I (self-consciously) look outward.' But what makes this so special? The reason that it is distinguished from other kinds of self-knowledge (perhaps arrived at through testimony), and from knowledge of another person's mental attitudes, is that the knowledge doesn't function as just mere information. Rather, transparent self-knowledge puts me in a position 'to speak of [my] conviction about the facts' (Moran 2001, 76). In a similar vein, if I am aware of a particular intention of mine, this awareness doesn't involve a *predictive* attitude toward my future action, but rather 'my seeing a certain act as in my power and regarding it as the thing to do' (Boyle 2019, 12). For instance, if I avow my intention to watch a Wes Anderson movie tonight, this has less to do with "getting the facts right" as with *leading my life*. Where another person might roll their eyes that I will watch a Wes Anderson movie *again* and go on with their day, this is not how I can relate to my intention. If this is my intention, then this calls on me to exercise my power to make it happen.

What do such considerations tell us about transparent self-knowledge? Such considerations do not seem to make transparent self-knowledge different from other kinds of self-knowledge or knowledge of other's mental attitudes in *what* it is; they all concern facts about someone having a particular attitude. Instead, it seems that they are considerations about *the way in which* transparent self-knowledge exists: the logical form of self-knowledge. Future research of transparent self-knowledge should therefore, in my view, elaborate on the idea that the distinctiveness of transparent self-knowledge lies in its logical form. Such an inquiry would require an analytic Aristotelian approach.

3. Taking a broader perspective

It is obvious that the kind of analysis in this dissertation is at odds with the scientistic perspective of self-knowledge presented in the introduction. While scientism as such hasn't been explicitly considered in any of the chapters, it has been playing a role in the background. Let me mention three ways in which scientism has manifested itself.

First of all, the account of transparent self-knowledge discussed in the dissertation that seems to be the most compatible with contemporary views in psychology is the reliabilist construal of Byrne's inferentialist account (Chapter 2).

After all, if a psychological mechanism is to produce knowledge, it should be reliable. In this view, the subject has transparent self-knowledge if she reliably infers that she believes that *p* from the premise *p*. She does so reliably if she actually believes that *p*. As we have seen, however, the reliabilist construal has unwelcome consequences. For it implies that (as argued in Chapter 2), in order not to presuppose self-knowledge of one's belief that *p* from the outset, the subject not only doesn't need to be aware of the procedure but cannot become aware of it. If she were, she wouldn't be in a position to endorse its cogency (it would still be a "mad inference"). This in turn implies that the subject is not in a position to connect her self-ascription of the belief that *p* to her actual views on *p*. To avoid such absurd consequences, the reliabilist construal should be rejected. If the reliabilist construal is indeed in line with contemporary views in psychology, this gives us reason to doubt that contemporary views in psychology are in a position to adequately account for transparent self-knowledge.

Secondly, psychology and neuroscience seek to explain mental phenomena in terms of mental states and processes and the underlying mechanisms. This scientific perspective motivates a kind of reductionist analysis of mental phenomena in terms of smaller parts (often mental attitudes or states) and the relation between these parts (often causal explanatory relations). We see this for example in the debate on self-knowledge, reasoning, and intentional action. In the case of reasoning, we have seen that this means that reasoning is analyzed in terms of moving from premise-beliefs to a conclusion-belief (Chapter 3). As I have argued, this view of reasoning runs into many problems. In spite of this, there is little room in the debate to approach reasoning differently, i.e., without analyzing it in terms of smaller parts (the mental attitudes and the relations between them). There is little room, in other words, to resist the scientific perspective of the mind and develop, for instance, the form view of reasoning.

Finally, I have investigated the possibility that a person can infer her mental attitudes from her patterns of action and reaction (Chapter 5). Is it possible to ignore the subject's active relation to her own mental life and the agency manifested in taking and having a stance? What I have argued is that, even if some mental attitudes might be thus inferred, each instance of inferring depends on the subject taking a stance. Mental data, inner promptings, and patterns of action and reaction have symptomatic value only if, somewhere down the line, the subject takes them to be expressive of her perspective. Mental data and internal promptings aren't just *given facts* about the subject that she can discover, experience or note, but are themselves an expression of the subject's commitments to the world at large. They are, at least always partly, an expression of her agency. It thus seems that we cannot do without

the first-person agential perspective. This gives us reason to think that a wholly scientific third-personal perspective of the mind cannot be adequate.

I don't mean to be claiming that these considerations would convince those in favor of scientism nor those who disregard the first-person perspective. Nonetheless, if my arguments are sound, then the very practice of science itself depends on the non-reductive exposition of reasoning and self-knowledge advanced in this dissertation. It thus seems that science cannot bypass the agential perspective.

Nor am I claiming that my approach couldn't benefit from closer attention to the scientific results. This would be a worthwhile and exciting line of future research. The point I want to make is that a genuine understanding of the nature of our mental life requires us to reject a wholly mechanistic view of our minds. As Philippa Foot puts it in her *Moral Arguments* (1958, 509):

> …evidence is not a sort of medicine which is taken in the hope that it will work… When given good evidence, it is one's business to act on it, not to hang around waiting for the right state of mind…

Foot's remarks are spot on. Evidence, reasons, inner promptings, and mental attitudes require the subject to take a stance – to manifest her agency. Rather than waiting for the right state of mind, the reflections presented in this dissertation are an invitation to take a stance; an invitation to reconsider the importance of transparent self-knowledge and the indispensability of a thoroughgoing philosophical analysis of self-knowledge.

BIBLIOGRAPHY

Anscombe, G.E.M. 1957. *Intention*. Oxford: Basil Blackwell.

Anscombe, G.E.M. 1989. "Von Wright on Practical Inference." In *The Philosophy of Georg Henrik von Wright*, edited by Paul A. Schilpp and Lewis E. Hahn, 377-404. La Salle, Illinois: Open Court.

Armstrong, David M. 1968. *A Materialist Theory of the Mind*. London: Routledge.

Arpaly, Nomy. 2003. *Unprincipled Virtue. An Inquiry into Moral Agency*. Oxford: Oxford University Press.

Ashwell, Lauren. 2013a. "Deep. Dark,… or Transparent? Knowing Our Desires." *Philosophical Studies* 165 (1): 245-256.

Ashwell, Lauren. 2013b. "Review of Transparent Minds: A Study of Self-Knowledge, by Jordi Fernández." *Notre Dame Philosophical Reviews*, 8.

Audi, Robert N. 1998. *Epistemology*. London: Routledge.

Barnett, David J. 2016. "Inferential Justification and the Transparency of Belief." *Nous* 50 (1): 184-212.

Bayne, Tim and Michelle Montague (eds.). 2011. *Cognitive Phenomenology*. Oxford: Oxford University Press.

Bilgrami, Akeel. 2006. *Self-knowledge and Resentment*. Cambridge, Mass.: Harvard University Press.

Boghossian, Paul. 1989. "Content and Self-Knowledge." *Philosophical Topics* 17 (1): 5-26.

Boghossian, Paul. 2014. "What is inference?" *Philosophical Studies* 169: 1-18.

Boyle, Matthew. 2005. *Kant and the Significance of Self-Consciousness*. PhD Dissertation.

Boyle, Matthew. 2009a. "Active Belief." *Canadian Journal of Philosophy* 39 (suppl. 35): 119-147.

Boyle, Matthew. 2009b. "Two Kinds of Self-Knowledge." *Philosophy and Phenomenological Research* 78 (1): 133-164.

Boyle, Matthew. 2011a. ""Making up your Mind" and the Activity of Reason." *Philosopher's Imprint* 11 (17): 1-24.

Boyle, Matthew. 2011b. "Transparent Self-Knowledge." *Aristotelian Society* 85 (suppl.): 223-241.

Boyle, Matthew. 2015. "Critical Study: Cassam on Self-Knowledge for Humans," *European Journal of Philosophy* 23 (2): 337-348.

Boyle, Matthew. 2019. "Transparency and reflection." *Canadian Journal of Philosophy* (forthcoming): 1-28.

Broome, John. 2013. *Rationality through Reasoning*. Sussex: Blackwell.

Broome, John. 2014. "Comments on Boghossian." *Philosophical Studies* 169: 19-25.

Burge, Tyler. 1998. "Reason and the first person." In *Knowing Our Own Minds*, edited by Crispin Wright, Barry C. Smith and Cynthia Macdonald, 243-270. Oxford: Oxford University Press.

Byrne, Alex. 2005. "Introspection." *Philosophical Topics* 33: 79-104.

Byrne, Alex. 2011. "Transparency, Belief, Intention." *Aristotelian Society* 85 (supp.): 201-221.

Byrne, Alex. 2015. "Skepticism about the Internal World." In *The Norton Introduction to Philosophy*, edited by Gideon Rosen, Alex Byrne, Joshua Cohen, and Seana V. Shiffrin. New York, NY: W. W. Norton & Company, Inc.

Byrne, Alex. 2018. *Transparency and Self-Knowledge*. Oxford: Oxford University Press.

Carroll, Lewis. 1895. "What the Tortoise Said to Achilles." *Mind* 4 (14): 278-280.

Carruthers, Peter. 2011. *The Opacity of mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Cassam, Quassim. 2011. "How We Know What We Think." *Revue de Métaphysique Et de Morale* 4 (4): 553–569.

Cassam, Quassim. 2014. *Self-knowledge for Humans.* Oxford: Oxford University Press.

Chisholm, Roderick M. 1982. *The Foundations of Knowing*. Minneapolis: University of Minnesota Press.

Coliva, Annalisa. 2016. *The Varieties of Self-Knowledge*. London: Palgrave Macmillan UK.

Coliva, Annalisa. 2014. "Review of Jordi Fernández *Transparent Minds*." *Theoria* 81: 442-445.

Dancy, Johathan. 1995. "Arguments from illusion." *Philosophical Quarterly* 45 (181): 421-438.

Dennett, Daniel C. 1978. "Conditions of Personhood." In *Brainstorms*: 267-285. Boston, MA: MIT Press.

Dennett, Daniel C. 1987. *The Intentional Stance*. Boston, MA: MIT Press

Dennett, Daniel C. 2017. *From Bacteria to Bach and Back: The Evolution of Minds*. New York: W.W. Norton & Company.

De Sousa, Ronald. 2007. "Truth, Authenticity, and Rationality." *Dialectica* 61 (3): 323–345.

Döring, Sabine A. 2007. "Seeing What to Do: Affective Perception and Rational Motivation." *Dialectica* 61 (3): 363–394.

Doris, John M. *Talking to Ourselves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press.

Doyle, Casey. (2018). "Agency and observation in knowledge of one's own thinking." *European Journal of Philosophy*: 1-14.

Dretske, Fred. 2003. "How Do You Know You Are Not a Zombie?" In *Privileged Access: Philosophical Accounts of Self-Knowledge*, edited by Brie Gertler. Burlington: Ashgate Publishing Company.

Edgley, Roy. 1969. *Reason in Theory and Practice*. London: Hutchinson.

Evans, Gareth. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Falvey, Kevin. 2000. "Knowledge in Intention." *Philosophical Studies* 99 (1): 21-44.

Fernández, Jordi. 2013. *Transparent Minds: A Study of Self-Knowledge*. Oxford: Oxford University Press.

Finkelstein, David. 2003. *Expression and the Inner*. Cambridge MA: Harvard University Press.

Finkelstein, David. 2012. "From Transparency to Expression." In *Rethinking Epistemology*, *vol. 2*, edited by Gunter Abel and James Conant: 101-118. Berlin/Boston: De Gruyter.

Foley, Richard. 2001. *Intellectual Trust in Oneself and Others*. New York: Cambridge University Press.

Foot, Philippa. 1958. "Moral Arguments." *Mind* 67 (268): 502-513.

Ford, Anton. 2015. "The Arithmetic of Intention." *American Philosophical Quarterly* 52 (2):129-143.

Frege, Gottlob. 1979. "Logic." In *Posthumous Writings*: 1-8. Chicago: University of Chicago Press.

Frey, Jennifer. 2013. "Analytic philosophy of action: a very brief history." *Philosophical News* 7: 50-58.

Gallois, Andre. 1996. *The World Without, the Mind Within: An Essay on First-Personal Authority*. Cambridge: Cambridge University Press.

Geach, Peter. 1957. *Mental Acts*. London: Routledge & Kegan Paul Ltd.

Gelfert, Axel. 2014. *A Critical Introduction to Testimony*. London: Bloomsbury.

Gertler, Brie. 2011. "Self-Knowledge and the Transparency of Belief." In *Self-Knowledge*, edited by Anthony Hatzimoysis: 125-145. Oxford: Oxford University Press.

Gertler, Brie. 2015. "Self-knowledge." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Gertler, Brie. 2016. "Critical notice of Quassim Cassam, *Self-Knowledge for Humans*." *Mind* 125: 269-280.

Goldie, Peter. 2002. "Emotions, Feelings, and Intentionality." *Phenomenology and the Cognitive Sciences* 1: 235-254.

Goldman, Alvin I. 1967. "A Causal Theory of Knowing." *The Journal of Philosophy* 64 (12): 357-372.

Goldman, Alvin I. 1993. "The Psychology of Folk Psychology." *Behavioral and Brain Sciences* 16: 15-28.

Goldman, Alvin I. 2006. *Stimulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Haldane, John. 2012. "Scientism and its Challenge to Humanism." *New Blackfriars* 93 (1048): 671-686.

Hampshire, Stuart. 1975 [1965]. *Freedom of the Individual*. Princeton: Princeton University Press.

Harman, Gilbert. 1986. *Change in View: Principles of Reasoning*. Cambridge, Mass.: MIT Press.

Harman, Gilbert. 1990. "The intrinsic quality of experience." *Philosophical Perspectives* 4: 31-52.

Heal, Jane. 2004. "Moran's 'Authority and Estrangement'." *Philosophy and Phenomenological Research* 69 (2): 427–432.

Helm, Bennett W. 2009. "Emotions as Evaluative Feelings." *Emotion Review* 1 (3): 248–255.

Helm, Bennett W. 2010. "Emotions and Motivation: Reconsidering Neo-Jamesian Accounts." In *The Oxford Handbook of Philosophy of Emotion*, edited by Peter Goldie: 303-324. New York: Oxford University Press.

Hieronymi, Pamela. 2009. "Two Kinds of Agency." In *Mental Actions and Agency*, edited by Lucy O'Brien and Matthew Soteriou: 138-162. Oxford: Oxford University Press.

Hlobil, Ulf. 2014. "Against Boghossian, Wright and Broome on Inference." *Philosophical Studies* 167 (2): 419-429.

Hlobil, Ulf and Katharina Nieswandt. 2016. "On Anscombe's Philosophical Method." *Klēsis Revue Philosophique* 35: 180-198.

Hofmann, Frank (2018). "How to know one's experiences transparently?" *Philosophical Studies*: 1-20.

Jones, Karen. 2004. "Emotional Rationality as Practical Rationality." In *Setting the Moral Compass: Essays by Women Philosophers*, edited by Cheshire Calhoun: 333-352. New York: Oxford University Press.

Jongepier, Fleur. 2017. *The Circumstances of Self-Knowledge*. PhD Dissertation.

Kind, Amy. 2003. "What's so transparent about transparency?" *Philosophical Studies* 115: 225-244.

Kloosterboer, Naomi. 2015. "Transparent Emotions? A Critical Analysis of Moran's Transparency Claim." *Philosophical Explorations* 18 (2): 246-258.

Knorpp, William M. 1997. "The Relevance of Logic to Reasoning and Belief Revision: Harman on 'Change in View'." *Pacific Philosophical Quarterly* 78: 78-92.

Kompridis, Nikolas. 2000. "So We Need Something Else for Reason to Mean." *International Journal of Philosophical Studies* 8 (3): 271-295.

Korsgaard, Christine M. 2009. *Self-Constitution: Agency, Identity and Integrity*. Oxford: Oxford University Press.

Korsgaard, Christine M. 1996. *The Sources of Normativity*. Edited by Onora O'Neill. Cambridge: Cambridge Universtiy Press.

Ladyman, James. 2011. "The Scientistic Stance: The Empirical and Materialist Stances Reconciled." *Synthese* 178 (1): 87-98.

Lamme, Victor. 2011. *De Vrije Wil Bestaat Niet*. Amsterdam: Bert Bakker.

Lawlor, Krista. 2009. "Knowing What One Wants". *Philosophy and Phenomenological Research* 79 (1): 47-75.

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

Maier, John. 2013. "The Agentive Modalities." *Philosophy and Phenomenological Research* 90 (1): 113-134.

Martin, M.G.F. 1998. "An Eye Directed Outward." In *Knowing Our Own Minds*, edited by Crispin Wright, Barry C. Smith and Cynthia Macdonald, 99-122. Oxford: Oxford University Press.

McGeer, Victoria. 1996. "Is 'Self-Knowledge' an Empirical Problem? Renegotiating the Space of Philosophical Explanation." *Journal of Philosophy* 93 (10): 483-515.

McGeer, Victoria. 2007. "The moral development of First-Person Authority." *European Journal of Philosophy* 16 (1): 81-108.

McHugh, Conor and Jonathan Way. 2016. "Against the Taking Condition." *Philosophical Issues* 26 (1): 314-331.

McHugh, Conor and Jonathan Way. 2018. "What is Reasoning?" *Mind* 127 (505): 167-196.

Moore, G.E. 1903. "The Refutation of Idealism." *Mind* 12 (48), 433-453.

Moore, G.E. 1993. "Moore's Paradox." In *G.E. Moore: Selected Writings*, edited by Thomas Baldwin, 207-212. London: Routledge.

Moran, Richard. 2001. *Authority and Estrangement. An Essay on Self-Knowledge*. Princeton & Oxford: Princeton University Press.

Moran, Richard. 2003. "Responses to Shoemaker and O'Brien." *European Journal of Philosophy* 11: 402-419.

Moran, Richard. 2004a. "Replies to Heal, Reginster, Wilson, and Lear." *Philosophy and Phenomenological Research* 69 (2): 455-72.

Moran, Richard. 2004b. "Anscombe on 'Practical Knowledge'." In *Royal Institute of Philosophy Supplement*, edited by John Hyman and Helen Steward: 43-68. Cambridge: Cambridge University Press.

Moran, Richard. 2008. "Frankfurt on Identification: Ambiguity of Activity in Mental Life." In *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton: 189-217. Cambridge, Mass.: MIT Press.

Moran, Richard. 2012. "Self-Knowledge, 'Transparency', and the Forms of Activity." In *Introspection and Consciousness*, edited by Declan Smithies and Daniel Stoljar: 211-235. Oxford: Oxford University Press.

Müller, A.W. 1979. "How theoretical is practical reason?" In *Intention and Intentionality: Essays in Honour of G. E. M. Anscombe*, edited by Cora Diamond and Jenny Teichman: 91-108. Sussex: Harvester Press.

Nagel, Thomas. 1996. "Universality and The Reflective Self." In *Sources of Normativity*, edited by Onora O'Neil: 200-209. Cambridge: Cambridge University Press.

Nichols, Shaun and Stephen P. Stich. 2003. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Oxford University Press.

Nida-Rümelin, Martine. 2007. "Transparency of Experience and the Perceptual Model of Phenomenal Awareness." *Philosophical Perspectives* 21: 429-455.

Nisbett, Richard and Timothy Wilson. 1977. "Telling more than we can know: Verbal reports on mental processes." *Psychological Review* 84: 231-259.

Nozick, Robert. 1981. *Philosophical Explanations*. Cambridge, Mass.: Harvard University Press.

O'Brien, Lucy. 2003. "Moran on Agency and Self-Knowledge." *European Journal of Philosophy* 11 (3): 391-401.

O'Brien, Lucy. 2007. *Self-Knowing Agents*. Oxford: Clarendon Press.

Ometto, Dawa. 2016. *Freedom and Self-Knowledge*. PhD Dissertation.

Paul, Sarah K. 2012. "How we know what we intend." *Philosophical Studies* 161: 327-346.

Paul, Sarah K. 2014. "The Transparency of Mind." *Philosophy Compass* 9 (5): 295-303.

Peacocke, Antonia. 2017. "Embedded mental action in self-attribution of belief." *Philosophical Studies* 174 (2): 353-377.

Peacocke, Christopher. 1998. "Conscious Attitudes, Attention, and Self-Knowledge." In *Knowing Our Own Minds*, edited by Crispin Wright, Barry C. Smith and Cynthia Macdonald, 63-98. Oxford: Oxford University Press.

Peacocke, Christopher. 1999. *Being Known*. Oxford: Oxford University Press.

Peacocke, Christopher. 2007. "Mental Action and Self-Awareness." In *Contemporary Debates in the Philosophy of Mind*, edited by Jonathan D. Cohen and Brian P. McLaughlin: 358-375. Oxford: Blackwell.

Peacocke, Christopher. 2009. "Mental Action and Self-Awareness (II): Epistemology." In *Mental Actions*, edited by Lucy O'Brien and Matthew Soteriou: 193-215. Oxford: Oxford University Press.

Peels, Rik. 2018. "A Conceptual Map of Scientism." In *Scientism: Prospects and Problems*, edited by Jeroen D. de Ridder, Rik Peels, and René van Woudenberg. Oxford: Oxford University Press.

Pippin, Robert B. 2005. "On 'Becoming Who One Is' (and Failing): Proust's Problematic Selves." In *The Persistence of Subjectivity: on the Kantian Aftermath*: 307-338. New York: Cambridge University Press.

Plantinga, Alvin. 1993. *Warrant and Proper Function*. New York: Oxford University Press.

Pritchard, Duncan. 2005. *Epistemic Luck.* Oxford: Oxford University Press.

Reed, Baron and Diego Machuca (eds.). 2018. *Skepticism: From Antiquity to the Present*. London: Bloomsbury.

Reid, Thomas. 1764 [1997]. *Inquiry into the Human Mind on the Principles of Common Sense*. Edited by Derek Brookes. Edinburgh: Edinburgh University Press.

Ridder, Jeroen D. de, Rik Peels, and René van Woudenberg. 2018. *Scientism: Prospects and Problems*. Oxford: Oxford University Press.

Rödl, Sebastian. 2007. *Self-consciousness*. Cambridge, Mass.: Harvard University Press.

Roessler, Johannes. 2013a. "The Silence of Self-Knowledge." *Philosophical Explorations* 16 (1): 1-17.

Roessler, Johannes. 2013b. "The Epistemic Role of Intentions." *Proceedings of the Aristotelian Society* 113 (1): 41-56.

Rosenberg, Alex. 2011. *The Atheist's Guide to Reality: Enjoying Life without Illusions.* New York: W. W. Norton.

Rosenthal, D.M. 2005. *Consciousness and Mind*. Oxford: Clarendon Press.

Schöttler, Peter. 2012. "Szientismus. Zur Geschichte eines schwierigen Begriffs." *NTM Zeitschrift für Geschichte der Wissenschaften, Technik und Medizin* 20 (4): 245-269.

Schroeder, Tim. 2015. "Desire." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Schwenkler, John. 2018. "Self-Knowledge and its Limits." *Journal of Moral Philosophy* 15 (1): 85-95.

Schwitzgebel, Eric. 2010. "Acting Contrary to Our Professed Beliefs or the Gulf Between Occurrent Judgment and Dispositional Belief." *Pacific Philosophical Quarterly* 91 (4): 531-553.

Schwitzgebel, Eric. 2012. "Self-Ignorance." In *Consciousness and the Self*, edited by Jeeloo Liu and John Perry. Cambridge University Press.

Schwitzgebel, Eric. 2015. "Belief." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Seidman, Jeffrey. 2016. "The Unity of Caring and the Rationality of Emotion." *Philosophical Studies* 173 (10): 2785-2801.

Setiya, Kieran. 2011. "Knowledge of Intention." In *Essays on Anscombe's Intention*, edited by Anton Ford, Jennifer Hornsby and F. Stoutland: 170-197. Cambridge, Mass.: Harvard University Press.

Shah, Nisha and David Velleman. 2005. "Doxastic Deliberation." *The Philosophical Review* 114: 497-534.

Shoemaker, Sydney. 1996. *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.

Shoemaker, Sydney. 2003. "Moran on Self-Knowledge." *European Journal of Philosophy* 11 (3): 391-401.

Shoemaker, Sydney. 2009. "Self-Intimation and Second-Order Belief." *Erkenntnis* 71 (1): 35-51.

Silins, Nicholas. 2012. "Judgment as a Guide to Belief." In *Introspection and Consciousness*, edited by Declan Smithies and Daniel Stoljar: 295-317. New York: Oxford University Press.

Smith, Angela M. 2005. "Responsibility for Attitudes: Activity and Passivity in Mental Life." *Ethics* 115: 236-71.

Sosa, Ernest. 2009. *Reflective Knowledge: Apt Belief and Reflective Knowledge Volume II.* Oxford: Clarendon Press.

Soteriou, Matthew. 2009. "Mental Agency, Conscious Thinking, and Phenomenal Character." In *Mental Actions*, edited by Lucy O'Brien and Matthew Soteriou: 232-252. Oxford: Oxford University Press.

Spener, Maya. 2011. "Disagreement about Cognitive Phenomenology." In *Cognitive Phenomenology*, edited by Tim Bayne and Michelle Montague: 268-284. Oxford: Oxford University Press.

Stoljar, Daniel. 2004. "The Argument from Diaphanousness." In *New Essays in the Philosophy of Language and Mind*, edited by Maite Ezcurdia, Robert Stainton, Christopher Viger: 341-390. Calgary: University of Calgary Press.

Stroud, Barry. 2002. *Understanding Human Knowledge*. Oxford: Oxford University Press.

Stroud, Sarah. 2003. "Weakness of Will and Practical Judgement." In *Weakness of Will and Practical Irrationality*, edited by Sarah Stroud and Christine Tappolet: 121-146. Oxford: Oxford University Press.

Teroni, Fabrice. 2007. "Emotions and Formal Objects." *Dialectica* 61 (3): 395-415.

Thompson, Michael. 2008. *Life and Action: Elementary Structures of Practice and Practical Thought*. Cambridge: Harvard University Press.

Tugendhat, Ernst. 1986. *Self-consciousness and self-determination*. Transl. by Paul Stern. Cambridge, Mass./ London: MIT Press.

Tye, Michael. 1995. *Ten problems of consciousness*. Cambridge, Mass.: MIT Press.

Valaris, Markos. 2011. "Transparency as Inference: Reply to Alex Byrne." *Proceedings of the Aristotelian Society* 111 (2): 319-324.

Valaris, Markos 2014. "Reasoning and Regress." *Mind* 123 (489): 101-127.

Valaris, Markos. 2016. "Supposition and blindness." *Mind* 125 (499): 895-901.

Valaris, Markos. 2017. "What Reasoning might be." *Synthese* 194: 2007-2024.

Valaris, Markos. 2018. "Reasoning and Deducing." *Mind*: 1-25.

Velleman, David. 2000. *The Possibility of Practical Reason.* Oxford: Oxford University Press.

Vogler, Candice. 2001. "Anscombe on Practical Inference." In *Varieties of Practical Reasoning*, edited by Elijah Millgram. Cambridge: MIT University Press.

Wilson, Timothy. 2002. *Strangers to Ourselves: Discovering the Adaptive Unconsciousness*. Cambridge, Mass.: Belknap Press.

Wittgenstein, Ludwig. 2009 [1953]. *Philosophical Investigations*. Transl. by G.E.M Anscombe. New York: Wiley.

Woudenberg, René van. 2005. "Intuitive Knowledge Reconsidered". In *Basic Belief and Basic Knowledge*, edited by René van Woudenberg, Sabine Roeser, Ron Rood: 15-39. Frankfurt: Ontos.

Woudenberg, René van, Jeroen de Ridder, Rik Peels. 2018. "Introduction." In *Scientism: Prospects and Problems*, edited by Jeroen D. de Ridder, Rik Peels, and René van Woudenberg. Oxford: Oxford University Press.

Woudenberg, René van, and Naomi Kloosterboer. 2019 (forthcoming). "Three Transparency Principles Examined." *Journal of Philosophical Research*.

Wright, Crispin. 2014. "Comment on Paul Boghossian, 'What is inference?'." *Philosophical Studies* 169: 27-37.

SUMMARY

Do I believe that it is raining? Do I believe that my partner and I will grow old together? Do I intend to pay back the money I borrowed? Do I prefer strawberries over raspberries? Do I value family over work? Should I focus on having fun, being a parent, a career woman, a good friend? These kinds of questions, both the more trivial and the more substantial ones, are central to this dissertation. They are the kinds of questions whose answers, if true, provide one with a piece of self-knowledge, namely, self-knowledge of one's own *intentional mental attitudes*.

Self-knowledge of mental attitudes is often regarded as special because such attitudes involve a specific first-person perspective, namely a commitment to what the attitude is about. For instance, if I believe that it is raining then I take it to be true that it is raining. And if I value family over work, I take family to have more importance than work. This connection between attitude and commitment, and especially the kind of self-knowledge that "respects" this connection, is the focus of this dissertation. I call this kind of self-knowledge *transparent self-knowledge.* It is transparent because it is a mode of knowing one's mental attitudes in which one, in a way to be specified, looks *beyond* or *through* the attitude to what the attitude is about. The underlying question addressed in this dissertation, which consists of five independent papers, is: how should transparent self-knowledge be conceptualized?

My study of transparent self-knowledge has been partly born out of amazement at the increasingly all-encompassing scientific perspective on the nature of the human mind. From the scientific perspective, all self-knowledge is regarded with skepticism and sometimes even declared illusory. To give an idea of this position, consider the following characterization of our capacity for self-knowledge by Daniel Dennett:

> …each of us is in most regards a sort of inveterate auto-psychologist, effortlessly *inventing* intentional interpretations of our own actions in an inseparable mix of confabulation, retrospective self-justification, and (on occasion, no doubt) good theorizing. (Dennett 1987, 91; emphasis in original)

Self-attributions of mental attitudes, in this view, are nothing more than the result of theorizing about what could go on in our minds that would explain what we do. Theorizing that is often better left to persons other than ourselves, for our own theorizing is obfuscated by our self-conception.

The departure point of this dissertation is to question the scientific skepticism about self-knowledge and its underlying assumptions about the nature of self-knowledge. What seems to be most problematic about the view of self-knowledge purported by science is the relation between a person and her own mental life that it presupposes. Their skepticism about self-knowledge addresses the idea of a person who must take on the role of a psychologist and who then tries to observe what is going on in her own mind. Like a bystander who merely witnesses what goes on in her head. As opposed to being a witness or bystander, a person doesn't merely *register* what is present in her mind. Rather, her mental goings-on express her view of things, i.e., what she takes to be true, what she will do, how she feels about things and what she wants. Her mental attitudes should thus be seen by her 'as expressive of [her] various and evolving relations to [her] environment, and not as a mere succession of representations (to which, for some reason, [she] is the only witness)' (Moran 2001, 32). Seeing one's attitudes as expressive of one's stance on the world at large implies that a person doesn't relate to her mental life as a psychologist but as someone who is actually inhabiting the perspective purported by those attitudes – as someone who is committed to the various things inherent in holding different mental attitudes. Given that transparent self-knowledge respects this connection between attitude and commitment, it can be seen as *the alternative* to the alienated scientific view of self-knowledge that has led to so much doubt and suspicion regarding the human capacity to know our own minds.

What is transparent self-knowledge?

Back to the underlying central question of the dissertation: how should transparent self-knowledge be conceived? First of all, let me relate it to philosophy of self-knowledge more generally. A principal point of departure in thinking about self-knowledge is the difference between knowledge of one's own mental attitudes and the mental attitudes of others. This difference is predominantly viewed in light of its epistemology: self-knowledge is thought to be *privileged*, i.e. to have a more secure epistemic status, and to be available through *peculiar access*, i.e. through means available only in knowing one's own mental attitudes (cf. Byrne 2005; Gertler 2015). What is often left out of these discussions of the epistemology of self-knowledge is the connection between self-knowledge and the nature of the person who is seeking

self-knowledge. Why should self-knowledge matter to us? What are the connections of self-knowledge to personhood, to moral psychology, and to (mental) agency? In what way do these moral psychological issues inform the difference between self-knowledge and knowledge of someone else's mental attitudes?

A philosopher who has refocused attention in the philosophical debate on self-knowledge to such moral psychological issues is Richard Moran (2001). The essential difference between knowledge of one's own mental attitudes and the mental attitudes of others is, according to Moran, not a difference in privileged or peculiar access but a difference in the way a person is involved in her own mental life. Different from the relation a person may have to someone else's mental life, her relation to her own mental attitudes is the aforementioned first-personal agential stance. She isn't some expert witness, but inhabits the perspective purported by her mental attitudes. This is to say that in believing that it rains she takes it to be true that it is raining; in feeling hurt she commits herself to the view that someone wronged her or something is hurtful; in intending to go to the new Wes Anderson movie tonight, she commits herself to making it her business to be there.

In respecting the relation between attitude and commitment, transparent self-knowledge seeks to do justice to this essential difference between self-knowledge and knowledge of someone else's mental attitudes. It is self-knowledge that, in the case of belief, not merely puts me in a position to report that I *have* a certain belief, but 'to speak of [my] conviction of the facts' (Moran 2001, 76). Similarly, in the case of intention, in being aware of my intention to go to the movie tonight, I am not aware of some *likelihood* that I will end up at the movie theater tonight. Rather, I am aware of having it made *my business* to be there. Hence, the idea is that transparent self-knowledge is the kind of self-knowledge that a person has from the stance of *agency*. That is, as someone who has an active relation to her own mental attitudes: both to whom it matters what her mental attitudes are and who is involved in what attitudes she holds.

The limits of transparency procedures: its scope

Having given a first characterization of transparent self-knowledge is not yet to give an *account* of transparent self-knowledge. This would minimally require addressing the question how such transparent self-knowledge is to be achieved. It is generally thought that achieving transparent self-knowledge should be explicated as going through a *transparency procedure*. Broadly speaking, such a procedure can be called transparent if it involves answering an inward-directed question about one's mental attitude by answering the relevant outward-directed question about the content of the attitude. For instance, one can achieve self-knowledge of one's belief that *p*, if

one answers the question whether one believes that *p* by answering the question whether *p*. One main objective of this dissertation has been to analyze such transparency procedures. What I have argued is that these procedures face two kinds of limits: limits in scope and in solving what I have dubbed the Two Topics Problem (TTP). Let me discuss both in turn, before coming back to the relation between a transparency procedure and transparent self-knowledge.

Are there limits in the scope of transparency procedures? In the dissertation, I have looked at the scope of Moran's transparency claim. **Chapter 1** begins by addressing the question what Moran's transparency claim precisely consists of. Before being able to evaluate the claim that a person is to answer the question whether she has a belief that *p* by answering the question whether *p* is true, we need to know what is required in order to answer that latter question. Moran's work supports three different requirements: 1) that there aren't any conditions on how to answer the question whether *p*; 2) that one should refer to reasons in favor of *p*; and 3) that one should refer to reasons justifying *p*. These three requirements are evaluated by checking whether each of them holds for numerous examples of belief, ranging from recalcitrant beliefs to beliefs based on no evidence. Take, for instance, Elisabeth's belief based on non-justifying reasons. Years after the war is over, Elisabeth keeps on believing that her husband will return home from said war. When asked why she believes this, she cites as her reason that there have been other men who have returned home after the war, although she is also aware of many countervailing reasons. Since her belief isn't based on justifying reasons, she cannot meet the third requirement: she cannot answer whether it is true that her husband will return home by reference to reasons justifying that proposition. However, the fact that her belief isn't based on justifying reasons doesn't preclude her having any reason in favor of the proposition believed. Hence, she might fulfill the second requirement. As such, she can also meet the first requirement, because that requirement attaches no conditions to how the question about the proposition is to be answered.

Based on expositions and considerations such as these, Moran's transparency claim seems most plausible, i.e. has the widest scope, if it means that the first requirement should be met. This isn't very surprising, of course, because the first requirement is fairly minimal. But because it is so minimal, it also has some counterintuitive results. It actually seems to be in tension with a person's active relation to her own mental life – to what Moran calls the *deliberative stance*. The exposition of the case of Elizabeth shows that she can meet the first requirement without taking a deliberative stance: she might stubbornly repeat the proposition believed without really being open to the question whether *to* believe that

proposition. This tension between the first requirement and the deliberative stance is also evidenced in another class of belief, namely obsessive beliefs. Even if a person only has reasons *against* believing that she will fail the exam, she might, when faced with the question 'Will I fail the exam?', cannot but think that she will. This means that she might meet the first requirement, although her relation to this belief bears no signs of the active relation that Moran seeks to incorporate in his account of self-knowledge. The upshot is that friends of Moran's transparency account either need to give up the deliberative stance (which was supposed to be its main motive) or justify its limited scope with respect to distinct classes of belief.

Another question about the scope of transparency procedures is whether they apply to mental attitudes other than belief, such as emotions, desires, cares, etcetera. **Chapter 4** takes up the question how Moran's account could be translated to mental attitudes other than belief. It assumes that Moran's transparency account works for belief and then seeks to apply it to emotion. The basic difficulty in such application is that where the relevant "outward-directed" question for belief is simply whether the proposition under consideration is true, the relevant "outward-directed" question for emotion is less easy to discern. The reason for this is that emotions do not only seem to be about the world, but also about what is important to the person having the emotion. Even if we all agree that a person has betrayed me, I need not *feel* betrayed if either the person or the betrayal itself were insignificant to me. Similarly, only if I care about a sports team, will their wins and losses spark joy or disappointment, respectively. We only feel an emotion if something matters to us (cf. Helm 2010). Chapter 4 thus argues that Moran's transparency claim cannot be applied to emotions, at least not without incorporating an account of the relation between transparency and what matters to us (a question that is addressed in Chapter 5).

Hence, Moran's delineation of a transparency procedure faces several limitations in its scope. First, it either fails to apply to several classes of belief or it becomes minimalistic to such a degree that it seems to be disconnected to a person's active relation to her mental attitudes. And secondly, it doesn't apply to emotions, nor, presumably, to other mental attitudes that are similarly connected to what matters to us.

The limits of transparency procedures: the two topics problem

This brings us to the second limitation of transparency procedures, namely the limit in solving TTP. The basic idea of transparency procedures is that a person answers a question about her own mental attitude by answering a relevant question about the topic of that mental attitude. The question is, however, why answering the latter

question is related to answering the former. What puts me in a position to answer the question whether I *believe* that it is raining by answering the questions whether it *is* raining? There seems to be something outright puzzling going on here, because the fact that it is raining doesn't entail that I believe that it is raining, nor provides evidence for it: one can imagine numerous scenarios in which it is raining but I do not believe that it is raining or in which I believe that it is raining but it isn't. Hence, transparency procedures face TTP: the problem that the truth of *p* doesn't seem to provide an epistemic basis for the truth of *I believe that p*.

A careful glance at the state of the debate on transparent self-knowledge shows that there is no consensus of what the relation between *p* and *I believe that p* might be, nor what kind of solution respects the commitments of transparency views that actually establish the source of TTP. **Chapter 2** seeks to provide a grasp on the nature of the different responses to TTP. The responses that I discuss are: 1) the view that TTP is only apparent; 2) inferential views; 3) judgment views; and 4) metaphysical views. In very general terms, the proceeding arguments are as follows. First, I argue that TTP has to be accepted as a genuine problem insofar as one accepts the transparency intuitions that in self-ascribing a belief that *p* a person both makes an empirical claim that she is in a certain state of mind and endorses *p*. Secondly, taking Alex Byrne's (2018) account as exemplary for the inferentialist response, I contend that a crucial assumption in his account, namely that inference from a premise entails belief in that premise, is unwarranted (a claim that is corroborated in Chapter 3). Thirdly, I argue that both the judgment views and the metaphysical views need, albeit for different reasons, to presuppose a form of attitudinal awareness, i.e. an awareness of one's judgment or belief regarding *p*. This is incompatible with delineating a *transparency* procedure to achieve self-knowledge. Hence, transparency procedures find their limit in TTP.

However, I don't think that this means that achieving transparent self-knowledge is impossible. We need to distinguish clearly between *transparent self-knowledge* and a *transparency procedure* to arrive at self-knowledge. Where the former expresses something about the nature of the self-knowledge at issue, the latter aims to identify an epistemically justified method of moving from thinking about the world at large to a self-ascription of a mental attitude. My explication of the nature of transparent self-knowledge as self-knowledge that manifests the connection between mental attitude and view of the world, doesn't imply that it necessarily depends on a transparency procedure. Even if the way in which one achieves transparent self-knowledge is not fully transparent, there is no immediate implication that there isn't another way in which such self-knowledge might be achieved. This is where I would like to point to the prospects of the involvement of

some form of attitudinal awareness, which was inherent in many of the accounts discussed in Chapter 2.

Since transparent self-knowledge is a mode of knowing in which one inhabits the perspective purported by the attitudes, invoking a form of attitudinal awareness to explain how we arrive at such self-knowledge brings with it a certain tension. After all, being aware of the attitudinal aspects of one's mental attitude implies that one isn't focused solely on the content of those attitudes. How could transparent self-knowledge be compatible with a form of attitudinal awareness? Such compatibility would seem to depend on the way in which this form of awareness modifies a person's relation to her own mental life. It shouldn't impugn on, most importantly, the relation between attitude and commitment. Put differently, we might say that the compatibility of transparent self-knowledge and the involvement of attitudinal awareness requires the latter to meet certain transparency requirements. For instance, the requirement that any uncertainty in one's self-ascription should direct one's attention back to the content of one's self-ascribed state (i.e. gazing back at the world). In my view, future research on this topic should thus focus on which transparency requirements attitudinal awareness should meet if it is to be compatible with transparent self-knowledge.

The form of self-knowledge

If transparency procedures face all these difficulties, one might wonder why again we needed transparent self-knowledge in the first place. The basic motivation for thinking there must be such a thing as transparent self-knowledge is to be able to make a genuine distinction between the first-person and third-person perspective (and thereby between self-knowledge and other-knowledge). What sets the first-person stance apart is, first and foremost, the connection between having a mental attitude and grasping the world in a certain way. What is especially important about transparent self-knowledge is thus that a person sees her mental attitudes not as part of the passing show but as expressive of her commitments. Seeing her attitudes as expressive of her own stance implies that the question of taking responsibility has become pertinent to all of them. When self-ascribing an attitude, we either take or *fail to take* responsibility for our attitudes by avowing or disavowing them. Avowing an attitude consists of a self-attribution of the attitude including an explicit endorsement of its content. In the case of belief, to avow my belief that *p* is to express my commitment to *p*'s truth. Because of this, achieving the kind of transparent self-knowledge at issue involves manifesting one's agency: for only if a person takes the responsibility to avow her mental attitude will she achieve transparent self-knowledge.

Safeguarding this stance of agency in matters of the human mind is thus, as argued in this dissertation, the reason why we need transparent self-knowledge. Nonetheless, different problems remain to be addressed. One worry is that a person's active relation to her mental attitudes is more present in the case of trivial attitudes (such as one's belief that it is raining) than in the case of substantial mental attitudes. Substantial mental attitudes, such as one's cares, concerns, deep desires and values, seem to be too integrated in one's entire life to depend merely on one's taking the responsibility to avow them. Were a person, say Katherine, to desire another child, this should be reflected not only in her avowal on the matter but also in a wide range of actions and reactions (cf. Lawlor 2009). It thus seems that transparent self-knowledge loses its relevance in the case of substantial attitudes.

In **Chapter 5** I take up this challenge and argue to the contrary: even if such patterns of action and reaction form part of coming to know my substantial mental attitudes, avowing these attitudes remains essential and has a unique status in coming to know them. My arguments show that the status of avowal is unique, first of all, because the significance of patterns of action and reaction, and what such patterns tell about our attitudes, ultimately depends on avowal. Secondly, they show that avowal is essential to knowing one's substantial mental attitudes, because these attitudes require one to have a self-conception. Acquiring self-knowledge of substantial mental attitudes can be seen as a struggle to fulfill the commitments pertaining to these attitudes. A struggle that requires a person to manifest her agency – to take responsibility for who she is and putting herself at risk of being challenged and making mistakes. In this view, avowal is, *par excellence*, essential to substantial self-knowledge, not only to trivial self-knowledge.

A final worry for providing a genuine conception of transparent self-knowledge is the wide variety to which it is supposed to apply. Following the literature, it seems as if we should slightly adjust the conception of transparent self-knowledge for each different mental attitude. How are we supposed to account for some kind of *unified* conception of transparent self-knowledge? The suggestion I want to make is that the analytic Aristotelian approach would be helpful to answer this question. I have introduced analytic Aristotelianism in **Chapter 3**, in a discussion on the nature of reasoning. Reasoning, like self-knowledge, comes in a wide variety. Chapter 3 develops an argument against the claim that *all* reasoning necessarily involves a change in attitudes (this argument also undermines the assumption that is required by Byrne's transparency procedure, as discussed in Chapter 2). Although it seems obvious that reasoning often involves such a change in attitudes, e.g., forming, revising or withdrawing a belief, that doesn't imply that a change in attitudes is necessarily involved in reasoning. For instance, we quite often

reason hypothetically or merely check the validity of an argument, without having determined for ourselves whether we believe the premises. As Wright (2014, 28) has put it, we should 'distinguish inference in general from *coming to a conclusion*…; no particular attitude to [a] proposition is implicit in inference itself.' By discussing examples of reasoning without a change in view, it will become clear that a different approach to reasoning is needed: namely, one that includes instances of reasoning with and without change in attitudes.

The alternative view that I develop is the *form view*. It holds that all the instances of reasoning can be unified if we adopt the view that reasoning is first and foremost *one thing*: namely recognizing the truth-connection between two statements (i.e. making a judgment that *p as following from q*). This is the *kind* of thing that reasoning is; its *logical form*. The logical form of a concept consists of the form of thought or the form of judgment that underlies the concept and it refers to what can be predicated of the thing in question, say X. The logical form (or *structure*) is revealed by analyzing the things that can be said or asked about X, and thus by analyzing our practices and abilities regarding X. I have argued that we need to understand reasoning in this way especially because of two reasons: first, it seems impossible to analyze reasoning in terms of sufficient and necessary conditions (or in terms of an explanatory essential feature or property), and secondly, because this seems to be the only way in which the extensive variety of instances of reasoning form a unity.

The suggestion I want to make is that something similar is true of transparent self-knowledge. I think that understanding the nature of transparent self-knowledge – understanding what it means to have self-knowledge of, say, one's belief that *p* in which one takes *p* to be true – requires an analytic Aristotelian approach. This accords with Moran's remark that

> …for a range of central cases, whatever knowledge of *oneself* may be, it is a very different thing from the knowledge of others, categorically different in kind and manner, different in consequences, and with its own distinguishing and constraining possibilities for success and failure. (Moran 2001, xxxi)

Future research of transparent self-knowledge should therefore, in my view, elaborate on the idea that the distinctiveness of transparent self-knowledge lies in its logical form. Such an inquiry would require an analytic Aristotelian approach.

Self-knowledge, science, and agency

I conclude this summary by returning to the opposition between a third-personal observational approach of self-knowledge that thrives in the scientific domain and a first-personal agential approach. The former approach rests on the assumption that mental attitudes can be inferred from a person's patterns of action and reaction (Chapter 5). But is it really possible to ignore the subject's active relation to her own mental life and the agency manifested in taking and having a stance? What I have argued is that, even if some mental attitudes might be thus inferred, each instance of inferring depends on the subject taking a stance. Mental data, inner promptings, and patterns of action and reaction have symptomatic value only if, somewhere down the line, the subject takes them to be expressive of her perspective. Mental data and internal promptings aren't just *given facts* about the subject that she can discover, experience or note, but are themselves an expression of the subject's commitments to the world at large. They are, at least always partly, an expression of her agency. It thus seems that we cannot do without the first-person agential perspective. This gives us reason to think that a wholly scientific third-personal perspective of the mind cannot be adequate.

By saying that science cannot bypass the agential perspective, I don't mean to claim that my approach couldn't benefit from closer attention to scientific results. This would be a worthwhile and exciting line of future research. The point I want to make is that a genuine understanding of the nature of our mental life requires us to reject a wholly mechanistic view of our minds. As Philippa Foot puts it in her *Moral Arguments* (1958, 509):

> …evidence is not a sort of medicine which is taken in the hope that it will work… When given good evidence, it is one's business to act on it, not to hang around waiting for the right state of mind…

Foot's remarks are spot on. Evidence, reasons, inner promptings, and mental attitudes require the subject to take a stance – to manifest her agency. Rather than waiting for the right state of mind, the reflections presented in this dissertation are an invitation to take a stance; an invitation to reconsider the importance of transparent self-knowledge and the indispensability of a thoroughgoing philosophical analysis of self-knowledge.

SAMENVATTING

Geloof ik dat het regent? Dat mijn partner en ik samen oud worden? Heb ik de intentie om het geleende geld terug te geven? Hou ik meer van aardbeien of van frambozen? Vind ik familie belangrijker of werk? Waarop moet ik me het meest richten: genieten van het leven, ouderschap, carrière, of vriendschap? Dit soort vragen, zowel de triviale als substantiële, staan centraal in dit proefschrift. De antwoorden op deze vragen bevatten, indien het wáre antwoorden zijn, zelfkennis. Zelfkennis van *intentionele mentale attitudes*, die begrepen kunnen worden als een geestestoestand die gericht is op iets. Voorbeelden zijn overtuigingen, verlangens, intenties en emoties.

Aan zelfkennis van mentale attitudes wordt vaak een speciale status toegekend, omdat zulke attitudes samenhangen met een eerste-persoonsperspectief. Overtuigingen, intenties, verlangens en emoties bestaan niet alleen "in mijn hoofd," maar zijn verbonden met mijn *commitment*, instemming met datgene waarover de attitude gaat. Als ik bijvoorbeeld geloof dat het regent, dan neem ik het voor waar aan dat het regent. En als ik meer geef om familie dan om werk, dan stem ik ermee in dat familie voor mij meer waarde heeft dan werk.

De verbinding tussen attitude en instemming, en in het bijzonder de zelfkennis die deze verbinding in acht neemt, behoort tot de kern van dit proefschrift. Deze vorm van zelfkennis noem ik *transparante zelfkennis*. Transparant omdat je jezelf in een bepaalde modus kent, waarbij je als het ware door je attitude heen kijkt naar datgene waar de attitude betrekking op heeft. De onderliggende vraag van dit proefschrift, bestaande uit vijf opzichzelfstaande artikelen, is dan ook: hoe kunnen we het concept van transparante zelfkennis systematisch analyseren en formuleren?

Mijn studie naar transparante zelfkennis is deels ingegeven door mijn verwondering over het feit dat het wetenschappelijke perspectief op de menselijke geest zich almaar uitbreidt. Vanuit dat wetenschappelijke perspectief wordt zelfkennis met scepsis bezien en soms zelfs tot illusie verklaard. Om een beeld te schetsen van deze positie kunnen we te rade gaan bij de Amerikaanse filosoof Daniel Dennett:

> …each of us is in most regards a sort of inveterate auto-psychologist, effortlessly *inventing* intentional interpretations of our own actions in an inseparable mix of confabulation, retrospective self-justification, and (on occasion, no doubt) good theorizing. (Dennett 1987, 91)

Ons vermogen tot zelfkennis wordt hier afgedaan als illusoir. We hebben geen toegang tot onze ware motieven. De verklaringen die we geven zijn niets meer dan het resultaat van interpretatie, waarbij we antwoord geven op de vraag wat er in ons hoofd om zou moeten gaan om ons gedrag afdoende te verklaren. Deze wetenschappelijk geïnspireerde visie op zelfkennis stelt zelfs dat, aangezien ons zelfbeeld onze eigen interpretaties troebleert, we deze interpretaties beter aan anderen kunnen overlaten. Die ander is een betere toeschouwer en psycholoog van onze mentale attitudes.

Dit proefschrift stelt vragen bij de wetenschappelijke scepsis over zelfkennis en dan met name bij het onderliggende beeld van zelfkennis, waarop deze scepsis is gebaseerd. Het meest problematisch aan het geschetste beeld van zelfkennis in de wetenschap is de relatie tussen een persoon en haar eigen mentale leven. Zij stellen een persoon voor die in de stoel van de psycholoog plaatsneemt en die, vanuit die positie, als een toeschouwer probeert te observeren wat zich in haar eigen hoofd afspeelt. Alsof zij alleen bezig is met de registratie van haar gedachten, gevoelens en attitudes. Maar een persoon is geen getuige van haar mentale leven. Haar mentale attitudes drukken uit hoe zij zich ten opzichte van de wereld verhoudt: wat zij als waar aanneemt, wat ze gaat doen, hoe ze zich voelt en waarnaar ze verlangt. Per slot van rekening is het haar mentale leven dat haar unieke standpunt in de wereld weergeeft. Om hier recht aan te doen, moet de persoon zelf daar besef van hebben: haar mentale attitudes moeten door haar gezien worden ‘as expressive of [her] various and evolving relations to [her] environment, and not as a mere succession of representations (to which, for some reason, [she] is the only witness)’ (Moran 2001, 32). Dit expressieve karakter van het eigen mentale leven houdt in dat een persoon zich niet als een psycholoog tot haar mentale leven verhoudt, maar als iemand die de attitudes als het ware “van binnenuit leeft.” Zo blijft de verbinding tussen attitude en commitment intact. Aangezien transparante zelfkennis deze verbinding in acht neemt, kan het gezien worden als hét alternatief voor het tot vervreemding leidende beeld van zelfkennis dat de wetenschap schetst en dat tot zoveel twijfel aan en achterdocht ten opzichte van zelfkennis heeft geleid.

Wat is transparante zelfkennis?

Laten we terugkeren naar de centrale vraag van dit proefschrift: hoe moeten we transparante zelfkennis begrijpen? Allereerst is het goed om dit vraagstuk in verband te brengen met de filosofie van zelfkennis in algemenere zin. Een fundamenteel startpunt in het denken over zelfkennis is het verschil tussen kennis van de eigen mentale attitudes en kennis van de attitudes van anderen. Meestal worden de epistemologische aspecten van dit verschil benadrukt: zo ziet men zelfkennis als *geprivilegieerd*, namelijk als kennis met een veilige epistemische basis, én als kennis die beschikbaar is via *unieke toegang*. Daarmee bedoel ik een toegang die alleen beschikbaar is wanneer het gaat om het eigen mentale leven (cf. Byrne 2005; Gertler 2015). Wat in epistemologische discussies over zelfkennis vaak buiten beschouwing blijft, is de relatie tussen zelfkennis en de aard van de naar zelfkennis strevende mens: waarom is zelfkennis belangrijk voor de mens?; hoe verhoudt zelfkennis zich tot ons persoonszijn, tot morele psychologie en tot (mentale) *agency*?; op welke manier werken deze moreel psychologische vraagstukken door in hoe we het verschil tussen zelfkennis en kennis van andermans attitudes moeten begrijpen?

Een filosoof die binnen het debat over zelfkennis de aandacht op deze moreel psychologische kwesties heeft gevestigd is Richard Moran. Het essentiële verschil tussen kennis van de eigen mentale attitudes en die van anderen ligt volgens Moran niet in *privilege* of *unieke toegang*, maar in de manier waarop een persoon betrokken is bij haar eigen mentale leven. Waar een persoon zich slechts als buitenstaander kan verhouden tot de mentale attitudes van anderen, beziet ze haar eigen mentale attitudes vanuit het eerdergenoemde eerste-persoonsperspectief. Zij is niet louter toeschouwer van wat er in haar hoofd gebeurt, maar ziet haar eigen perspectief op de wereld tot uitdrukking komen in haar eigen mentale attitudes. Waarmee ik bedoel dat ze, wanneer ze *gelooft* dat het regent, de waarheid van het feit *dat het regent* aanneemt; dat ze, wanneer ze zich gekwetst voelt, instemt met het idee dat iemand iets kwetsend gedaan heeft; en dat ze zich in het hebben van de intentie om vanavond de nieuwe Wes Anderson film in de bioscoop te bezoeken, erop toelegt dat daadwerkelijk te doen.

Door de verbinding tussen attitude en commitment in acht te nemen, tracht transparante zelfkennis recht te doen aan het daaraan verbonden verschil tussen zelfkennis en het kennen van andermans attitudes. Zo is het bij transparante zelfkennis van bijvoorbeeld een overtuiging het geval dat een persoon niet slechts in een positie is om te rapporteren dat zij een bepaalde overtuiging heeft, maar ook om haar instemming met datgene wat ze gelooft uit te spreken (cf. Moran 2001, 76). Bij zelfkennis van de intentie om vanavond naar de film te gaan, kan vergelijkbaar

gesteld worden dat het niet gaat om een bewustzijn van de waarschijnlijkheid dat de intentie verwezenlijkt wordt. Dat veronderstelt een tot vervreemding leidend perspectief van een buitenstaander, alsof het van een ander afhangt of iemand bij de bioscoop terechtkomt. In het kennen van een intentie gaat het er juist om dat een persoon zich ervan bewust is dat ze het tot haar doel gemaakt heeft om vanavond naar de bioscoop te gaan. Het idee is dus dat transparante zelfkennis het soort zelfkennis is vanuit het standpunt van een *agent*: iemand die zich actief tot haar mentale leven verhoudt.

De grenzen van transparantie-procedures

Met deze eerste kenschets van transparante zelfkennis heb ik nog geen theorie geformuleerd. Hiervoor zou op z'n minst een uitleg van de manier waarop zelfkennis vergaard wordt nodig zijn. In het algemeen stelt men dat dat het best begrepen kan worden als het doorlopen van een *transparantie-procedure*. De basis van zo'n procedure is bij benadering dat je een vraag over een mentale attitude beantwoordt ("Geloof ik dat het regent?") door een andere vraag te beantwoorden over datgene waar de attitude betrekking op heeft ("Regent het?"). Je gaat dus van de vraag "Geloof ik dat het regent?" naar de vraag "Regent het?" en gaat dan na of dit wel of niet het geval is. Zo is het idee dat een persoon transparante zelfkennis van haar geloof "dat *p*" (bijvoorbeeld geloof dat het regent) kan vergaren, als zij een antwoord geeft op de vraag of ze dat gelooft door de vraag "Is *p* waar?" te beantwoorden. Eén van de centrale oogmerken van dit proefschrift is het begrijpen en analyseren van dergelijke transparantie-procedures. Ik zie twee fundamentele problemen voor zulke procedures: grenzen aan het toepassingsgebied en aan de manier waarop zij gerechtvaardigd kunnen worden. Beide problemen zullen eerst kort besproken worden, voordat teruggekomen wordt op de verhouding tussen transparante zelfkennis en transparantie-procedures. Door deze verhouding te verhelderen, zien we dat transparante zelfkennis ook zonder transparantie-procedures in strikte zin kan blijven bestaan.

Zijn er grenzen aan de reikwijdte van transparantie-procedures? Deze vraag heb ik onderzocht door het bestek van Morans *transparantie-thesis* onder de loep te nemen. In **Hoofdstuk 1** begin ik met het expliciteren van zijn thesis om deze vervolgens te toetsen. De voorwaarden aan het beantwoorden van de vraag over datgene waarop de attitude betrekking heeft, dienen allereerst verhelderd te worden. Op geloof toegepast, gaat dit dus om de vraag "Is *p* waar?" (dus de vraag "Regent het?" in het bovengenoemde voorbeeld). Morans werk suggereert drie verschillende opties: 1) dat er geen condities aan de manier van beantwoorden gesteld worden; 2) dat gerefereerd moet worden aan redenen voor *p*; en 3) dat

gerefereerd moet worden aan redenen die geloof in *p* rechtvaardigen. Door deze drie voorwaarden op verschillende vormen van geloof toe te passen, worden ze geëvalueerd. Het resultaat van deze evaluatie is dat Morans transparantie-thesis de grootste reikwijdte heeft als aan de eerste en dus meest minimale voorwaarde voldaan moet worden. Maar omdat de voorwaarde zo minimaal is, lijkt zij losgezongen van de actieve relatie die een persoon ten opzichte van haar mentale leven heeft. En deze actieve relatie, het perspectief van een *agent*, was nu juist de drijvende kracht achter transparante zelfkennis. Gevolg hiervan is dat voorstanders van een Moran-achtige visie op transparantie-procedures moeten kiezen tussen twee uitdagingen: óf ze moeten rechtvaardigen dat transparante zelfkennis niet essentieel verbonden is met het perspectief van een *agent*, óf ze moeten de grenzen aan het bestek van de transparantie-procedure rechtvaardigen.

Een gerelateerde vraag over de reikwijdte van transparantie-procedures betreft de toepassing op andere attitudes dan geloof, zoals emoties, verlangens, intenties, geven om, etc. **Hoofdstuk 4** richt zich op de vraag of en hoe Morans theorie vertaald zou kunnen worden naar emotie. Allereerst neem ik aan dat Morans theorie werkt voor geloof. Vervolgens onderzoek ik hoe de theorie vertaald zou moeten worden. Het moeilijke is dat de te beantwoorden vraag bij een geloof dat *p* eenvoudig "Is *p* waar?" is, terwijl dit voor emotie minder eenduidig is. De reden ligt in de aard van emoties. Deze zijn namelijk niet alleen gericht op de wereld, maar hebben ook betrekking op wat voor iemand belangrijk is. Zelfs als we allemaal accepteren dat een bepaalde persoon mij verraden heeft, hoef ik me nog niet verraden te voelen als bijvoorbeeld de persoon voor mij onbelangrijk is of als het verraad als zodanig me koud laat. Evenzo zullen de winst en het verlies van een sportteam me slechts raken wanneer ik bij dat team betrokken ben. In algemene zin kunnen we stellen dat een persoon alleen een emotie voelt als iets voor haar belangrijk is (cf. Helm 2010). Via deze analyse beargumenteer ik dat Morans transparantie-thesis niet toepasbaar is op emotie, tenminste niet zonder een uiteenzetting over de relatie tussen transparantie en wat voor iemand belangrijk is (een vraag die ik behandel in Hoofdstuk 5). Ook zal de claim niet toepasbaar zijn op andere mentale attitudes die op dezelfde wijze verbonden zijn met wat voor iemand belangrijk is.

Tot zover een korte bespreking van het beperkte bereik van transparantie-procedures. Een andere grens waar deze procedures mee kampen is de rechtvaardiging van de verkregen zelfkennis. Transparantie-procedures gaan ervan uit dat iemand de ene vraag over een mentale attitude kan beantwoorden door een andere vraag over de inhoud van de attitude te beantwoorden. Maar wat is dan de relatie tussen beide vragen die een dergelijke methode mogelijk maakt? Normaliter

zijn vragen gerelateerd door een logisch of bewijsmatig verband, maar dat is bij transparantie niet het geval. Het feit dat het regent behelst noch indiceert dat ik dat geloof: er zijn talloze situaties denkbaar waarin het wel regent, maar ik dat niet geloof, of waarin ik wel geloof dat het regent, maar dat niet zo is. Zodoende staan transparantie-procedures tegenover het Twee Subjecten Probleem (*Two Topics problem*: TTP): namelijk dat de waarheid van de propositie *p* geen epistemische basis lijkt te verschaffen voor de waarheid van de propositie *ik geloof dat p*.

Indien we het huidige debat over transparante zelfkennis bestuderen, zien we dat consensus over een oplossing voor TTP ver te zoeken is. **Hoofdstuk 2** biedt een overzicht van de verschillende benaderingen van TTP en probeert zo het probleem zelf, maar vooral ook de relatie tussen transparantie-procedures en transparante zelfkennis te verhelderen. In het hoofdstuk beargumenteer ik dat transparante zelfkennis onlosmakelijk verbonden is met het idee dat het aan jezelf toeschrijven van een geloof dat *p* twee beweringen inhoudt: een empirische bewering dat de geest in die toestand verkeert, maar ook instemming met de waarheid van *p*. Met dit idee als uitgangspunt blijken alle oplossingen ontoereikend. Uiteindelijk kan transparante zelfkennis niet zonder enige vorm van bewustzijn van een attitude tot stand komen – en is dus niet uitsluitend gebaseerd op een bewustzijn van datgene waarop de attitude betrekking heeft. Als een dergelijke vorm van bewustzijn een rol speelt in het vergaren van transparante zelfkennis, dan kunnen we niet meer spreken over een zuivere transparantie-procedure. Dan is het namelijk niet zo dat de vraag over de attitude beantwoord wordt door alleen een antwoord te geven op de vraag over de inhoud van de attitude. TTP stelt dus een tweede grens aan transparantie-procedures.

Betekent dit dat transparante zelfkennis onmogelijk is? Niet als we transparante zelfkennis en transparantie-procedures zorgvuldig uit elkaar houden. De eerstgenoemde betreft de aard van de bedoelde zelfkennis, terwijl de laatstgenoemde tot doel heeft om een methode te identificeren die de sprong tussen instemming en attitude epistemisch rechtvaardigt. Ook als we stellen dat transparante zelfkennis de verbinding tussen attitude en instemming incorporeert, zoals besproken in dit proefschrift, hoeft dit nog niet te betekenen dat dit uitsluitend tot stand kan komen door een dergelijke sprong tussen instemming en attitude. De aard van transparante zelfkennis maakt een transparantie-procedure niet noodzakelijk. Het is goed mogelijk dat het vergaren van transparante zelfkennis berust op het bovengenoemde bewustzijn van een attitude. Let wel, zo'n vorm van bewustzijn staat haaks op het idee van transparantie, waarbij een persoon zogezegd door haar attitude heen kijkt naar datgene waarop de attitude betrekking heeft (en dus niet naar de attitude zélf). De rol en de aard van een dergelijk bewustzijn van

een attitude zal nader uitgewerkt moeten worden, met name zogenoemde transparantie voorwaarden waaraan zo'n bewustzijn dient te voldoen.

De vorm van transparante zelfkennis

Het bespreken van de moeilijkheden omtrent transparante zelfkennis heeft wellicht de vraag opgeroepen waarom transparante zelfkennis zo belangrijk is. Deze vorm van zelfkennis is nodig om het eerste- en derde-persoonsperspectief (en tevens zelfkennis en kennis van andermans attitudes) adequaat van elkaar te onderscheiden. Wat eigen is aan het eerste-persoonsperspectief is de verbinding tussen iemands attitudes en diens kijk op de wereld. Met transparante zelfkennis beziet een persoon haar mentale attitudes niet van een afstand, maar van binnenuit, als datgene wat haar instemming uitdrukt. Deze verbinding is direct verbonden met de verantwoordelijkheid die zij voor deze attitudes draagt. Een persoon neemt immers, omdat attitudes samenhangen met de relevante instemming, verantwoordelijkheid door deze instemming te erkennen, maar faalt als ze de betekenis van haar instemming ontkent. In het geval van een geloof dat *p* neemt een persoon verantwoordelijkheid als zij haar instemming met de waarheid van *p* erkent, in gedachten of expliciet. Zo beargumenteer ik, dat het bereiken van transparante zelfkennis een uitdrukking van *agency* vergt. Alleen als een persoon verantwoordelijkheid draagt en instemming erkent, kan er sprake zijn van transparante zelfkennis.

Het behouden van het perspectief van de agent inzake de menselijke geest toont de noodzaak van transparante zelfkennis. Desalniettemin blijven er problemen. Een belangrijk punt betreffende de rol van *agency* is dat deze lijkt af te nemen naarmate de mentale attitudes betekenisvoller worden. Wij lijken veel invloed te kunnen uitoefenen op attitudes die relatief weinig betekenis hebben (zoals de overtuiging dat het nu regent). Substantiële mentale attitudes (zoals onze zorgen, waar we om geven, onze waarden en diepe verlangens) zijn echter zo diep in ons leven verankerd, dat het wel of niet erkennen van instemming weinig lijkt uit te maken. Stel dat een persoon een verlangen naar een kind koestert. Het zou kortzichtig zijn om hierin alleen haar instemming te betrekken. Veel belangrijker is dat dit verlangen zich manifesteert in haar handelen, denken en gevoel (cf. Lawlor 2009). Is dit juist, dan heeft dit grote gevolgen voor transparante zelfkennis. Het zou namelijk betekenen dat deze vorm van zelfkennis haar relevantie verliest naarmate mentale attitudes betekenisvoller zijn.

In **Hoofdstuk 5** ga ik in op dit probleem en beargumenteer ik het tegenovergestelde: een dergelijk verlangen moet zich wel manifesteren in iemands handelen, denken en voelen, maar het nemen van verantwoordelijkheid door

instemming blijft essentieel om zelfkennis van dit verlangen te vergaren. De manifestaties van het verlangen in iemands handelen, denken en voelen kunnen alleen als betekenisvol herkend worden als er sprake is van instemming. Ook blijkt instemming cruciaal in het kennen van een dergelijk verlangen, omdat het samenhangt met iemands zelfbeeld. Het koesteren van een verlangen naar een kind hangt samen met het beeld dat je van jezelf hebt als moeder en als iemand die het waardevol vindt om een kind op de wereld te zetten, daarvoor te zorgen en een band mee op te bouwen. Zo bezien wordt het ook meteen duidelijk dat een dergelijk verlangen bepaalde toewijding vereist.

Kun je zelfkennis van dat verlangen hebben zonder op enige manier blijk te geven van een dergelijke toewijding? Mijn argumentatie in dit hoofdstuk laat zien dat het hebben van zelfkennis van substantiële attitudes het beste gezien kan worden als een strijd: een strijd om met een bepaald zelfbeeld het leven te leiden en tegen de eisen van toewijding aan te lopen. Soms als iemand die daar makkelijk aan kan voldoen, maar vaker als iemand die in die toewijding zal falen. Deze vorm van zelfkennis valt, *par excellence*, onder transparante zelfkennis: handelen met een zelfbeeld is een uiting van agency.

Het laatste probleem met betrekking tot het formuleren van een concept van transparante zelfkennis dat ik hier expliciet wil bespreken, is de verscheidenheid waarin transparante zelfkennis voorkomt. Als we de literatuur over transparante zelfkennis geloven, dan moeten we ons concept van transparante zelfkennis aan iedere verschillende mentale attitude aanpassen. Maar hoe komen we dan tot een samenhangend beeld van transparante zelfkennis? Ik ben van mening dat de analytisch Aristoteliaanse benadering hiervoor nodig is. Deze benadering introduceer ik in **Hoofdstuk 3** om de aard van redeneren te verhelderen. Ons vermogen tot redeneren lijkt op ons vermogen tot zelfkennis, omdat ook zij op zeer gevarieerde manieren tot uitdrukking komt. Dit zorgt ervoor dat analyses van redeneren – en zelfkennis, maar daar kom ik zo op terug – waarbij men zich verlaat op het formuleren van noodzakelijke voorwaarden, niet kunnen werken. Er blijkt niet één voorwaarde te zijn waaraan alle vormen van redeneren voldoen. Diegene die aangedragen worden als noodzakelijk, blijken niet voor alle gevallen te gelden (en dus niet noodzakelijk te zijn) of blijken niets anders te zijn dan een andere naam voor hetzelfde concept.

Hoofdstuk 3 laat dit zien door de onjuistheid aan te tonen van de bewering dat een verandering van mentale attitudes een noodzakelijke voorwaarde is van redeneren. Deze bewering stelt dat redeneren vereist dat je verandert van attitude. Hoewel het voor zich spreekt dat redeneren vaak een verandering van attitude inhoudt – bijvoorbeeld het vormen, herzien of herroepen van een overtuiging –

betekent dat nog niet dat het altijd zo is. Zo redeneren we hypothetisch of alleen maar om de geldigheid van een argument na te gaan, zonder dat we ons vergewissen of we overtuigd zijn van de premissen in het argument. Om hieraan recht te doen, hebben we een breder concept van redeneren nodig, dat geldt voor gevallen van redeneren met én zonder verandering van attitude.

Het ontwikkelde alternatieve beeld van redeneren richt zich niet op wát redeneren is, maar op "hoe" het is. In de analytisch Aristoteliaanse benadering wordt dit de *logische vorm* genoemd, waarmee bedoeld wordt dat 1) redeneren altijd bestaat uit een gedachte (ook wel oordeel genoemd) en 2) deze gedachte altijd een bepaalde structuur heeft. Dit is erg abstract en vergt nadere filosofische uitwerking. Desalniettemin toont Hoofdstuk 3 dat wanneer we redeneren op deze manier analyseren, duidelijk wordt waarom het concept zoveel verschillende verschijningsvormen heeft, maar toch samenhangend kan zijn.

Ook lijkt iets soortgelijks te gelden voor ons vermogen tot transparante zelfkennis, die zich ook kenmerkt door verscheidenheid. Haar aard kan mijns inziens in een samenhangend concept vervat worden als we de analytisch Aristoteliaanse benadering hanteren. Dit sluit aan bij de opmerking van Moran dat zelfkennis *categorisch* verschilt van kennis van andermans mentale leven (cf. Moran 2001, xxxi). Het idee dat transparante zelfkennis zich onderscheidt door haar logische vorm is, mijns inziens, veelbelovend.

Zelfkennis, wetenschap en agency

Afsluitend wil ik terugkomen op de tegenstelling tussen de in het wetenschappelijk domein florerende derde-persoons kijk op zelfkennis en de eerste-persoons, *agency*-gerelateerde benadering. Die eerste visie veronderstelt dat het gegeven dat een persoon een bepaalde mentale attitude heeft, volgt uit het feit dat zich in haar handelen, denken en voelen bepaalde patronen ontvouwen (zie Hoofdstuk 5). Dit laat de actieve houding die een persoon aanneemt ten opzichte van haar eigen mentale leven buiten beschouwing. Is dit in het licht van de in dit proefschrift gepresenteerde argumenten plausibel? Ik toon aan dat, zelfs als het bestaan van sommige mentale attitudes op deze manier geconcludeerd kan worden, een dergelijke conclusie altijd afhangt van de instemming van de persoon. Mentale *data*, zielenroerselen en patronen van actie en reactie kunnen alleen symptomatische waarde hebben als ze op een bepaalde manier verbonden zijn met het perspectief van de persoon zelf. Dergelijke informatie en innerlijke "oprispingen" zijn over de persoon geen gegeven feiten die zij kan ontdekken of registreren. Zij zijn zélf een uiting van het unieke standpunt dat de persoon inneemt. Ook zij zijn verbonden met de instemming van de persoon en dus deels een uiting van haar agency. Hieruit volgt

dat we niet zonder het perspectief van de *agent* kunnen. Dat leidt er mijns inziens toe een compleet wetenschappelijk derde-persoonsperspectief van de geest te wantrouwen.

Door te stellen dat wetenschap het perspectief van de *agent* niet links kan laten liggen, beweer ik nog niet dat mijn benadering volledig onafhankelijk is van de wetenschap. Integendeel, ik denk dat de door mij ontwikkelde visie op transparante zelfkennis verreikt zou kunnen worden door nadere beschouwing van wetenschappelijk onderzoek. Dit zou waardevol vervolgonderzoek zijn. Wat ik wil betogen is veeleer, dat een zuiver begrip van de aard van ons mentale leven alleen verkregen kan worden als we de derde-persoonsbenadering van de geest afwijzen. Philippa Foot omschrijft het als volgt:

> …evidence is not a sort of medicine which is taken in the hope that it will work… When given good evidence, it is one's business to act on it, not to hang around waiting for the right state of mind… (1958, 509)

Foot slaat hier, denk ik, de spijker op de kop: bewijs, redenen, zielenroerselen, en mentale attitudes vereisen een stellingname van de persoon – zij moet haar *agency* tot uiting brengen. De in dit proefschrift gepresenteerde overwegingen zijn dan ook een uitnodiging. Niet om achterover te leunen en te wachten op de juiste geestestoestand, maar om stelling te nemen – om de waarde van transparante zelfkennis en de onmisbaarheid van een filosofische kijk op zelfkennis te heroverwegen.

CURRICULUM VITAE

Naomi Kloosterboer (Amsterdam, 1986) obtained an interdisciplinary bachelor in the Exact and Social Sciences at the University of Amsterdam (*Bèta-gamma*, 2009, *cum laude*). She majored in philosophy and wrote a thesis on the neurophilosophy of free will. She did a research master's in philosophy at the same university, writing her thesis on the rationality of emotions (2012, *cum laude*). During her research master, she also spent a year at the Freie Universität and the Humboldt Universität in Berlin. Her dissertation, conducted at the Vrije Universiteit Amsterdam and part of the research project "Science beyond Scientism," was supervised by prof. dr. René van Woudenberg, prof. dr. Gerrit Glas, and dr. ing. Leon de Bruin. In 2015 and 2016, Naomi has been research fellow at Birmingham University and Harvard University. Currently, she is lecturer in practical philosophy at Utrecht University. Her research focuses on self-knowledge, agency and responsibility.