The Case for Pluralism about Introspection

By
Kateryna Samoilova
B.A., Harvard University, 2005

Dissertation
Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Philosophy at Brown University

Providence, Rhode Island
May 2013
This dissertation by Kateryna Samoilova is accepted in its present form by the Department of Philosophy as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date: 

Joshua Schechter, Advisor

Recommendation to the Graduate Council

Date: 

Richard Heck, Reader

Date: 

Christopher Hill, Reader

Date: 

Alex Byrne, Reader

Date: 

Susanna Siegel, Reader

Approved by the Graduate Council

Date: 

Peter Weber, Dean of the Graduate School
CURRICULUM VITAE

KATERyna SAMOILOVA

Education:
Brown University, Ph.D., Department of Philosophy, May 2013
Dissertation: The Case for Pluralism about Introspection
Committee: Joshua Schechter (principal advisor), Richard Heck, Christopher Hill, Susanna Siegel (Harvard), and Alex Byrne (MIT)
Harvard University, B.A., Cum Laude, Department of Philosophy, June 2005

Research Interests:
Areas of Specialization: Philosophy of Mind, Epistemology
Areas of Competence: Metaphysics, Philosophy of Psychology, Introductory Logic

Honors and Awards:
Roland G. D. Richardson Fund Fellowship, 2009-2010
Joukowsky Fellowship, 2006

Teaching Experience:
Visiting Lecturer, Department of Philosophy, Bridgewater State University
Spring 2008-present (9 semesters), Foundations of Logical Reasoning
Responsible for overall course design, including syllabus, assignments, and exams

Instructor, Department of Philosophy, Brown University
Summer 2012, Personal Identity
Fall 2008, Mind and Matter
Summer 2007, Personal Identity
Responsible for overall course design, including syllabus, assignments, and exams

Teaching Assistant, Department of Philosophy, Brown University
Spring 2007, Time, with Douglas Kutach
Fall 2006, Existentialism, with Bernard Reginster
Responsible for weekly discussion sections, office hours, and grading

Presentations:
Commentator, Nathan Ballantyne (Arizona), The 10th Annual Shapiro Graduate Philosophy Conference, Brown University, 2005

Co-organizer, 12th Annual Shapiro Graduate Philosophy Conference, 2007
Acknowledgements

I am grateful to my family, friends, and the members of the Philosophy Department at Brown University who have provided invaluable support and feedback over the years. I am particularly grateful to the faculty and the graduate students in the Dissertation Workshop and the Philosophy Graduate Forum at Brown. I would also like to thank the members of various readings groups at Harvard University, where I had an opportunity to present some parts of this dissertation. I am especially indebted to the members of my committee, Richard Heck, Christopher Hill, Alex Byrne, and Susanna Siegel, who have pushed me to consider the relevant issues from many different angles. And finally, I am deeply grateful to my principal advisor, Joshua Schechter, whose patience and enthusiasm have motivated me to constantly improve my work.
Table of Contents

Chapter 1: Introduction……………………………………………………………………1
Chapter 2: The Third Person Model of Introspection……………………………………26
Chapter 3: The Transparency Model of Introspection…………………………………43
Chapter 4: The Perceptual Model of Introspection………………………………………89
Chapter 5: A Complex Case of Introspection…………………………………………113
Chapter 6: Conclusion…………………………………………………………………..145
Bibliography……………………………………………………………………………151
Chapter 1: Introduction

Suppose you are now in the company of one of your friends, call him Sam, and he tells you that he’s been thinking about the current state of the economy and that this thought is giving him a headache. Suppose also that you know Sam is being sincere. Do you now know that Sam is thinking about the economy and that he has a headache? By everyday standards of what it takes to know other people’s mental states, you do. Nothing further needs to be settled about Sam before you can immediately and implicitly trust his judgment. We simply assume that a person’s access to one’s own current thoughts and headaches is epistemically secure, or at least authoritative.

How Sam managed to access his thoughts and his headache, what the epistemic security of this feat consists in, whether it is even epistemic, what it is a result of, and whether this security is not just illusory – these are the central issues surrounding the topic of introspection, the first-personal way of accessing our mental states that Sam intuitively relied on to gain a few bits of self-knowledge. I too plan to explore these issues in this dissertation, but my main goal is to defend a particular account of introspection, one that I call pluralism. Pluralism about introspection is the view that there is no single, unified way of introspectively gaining the body of self-knowledge we apparently possess. For example, even in Sam’s case, how he found out about his depressing thought and his headache require quite different explanations on my preferred pluralist account. As I present the arguments for this view, there are two motivating themes worth keeping in mind.
First, I want to expand the pool of cases we consider when talking about introspection. It is by no means easy analyzing even a simple-sounding, core case like Sam’s, which is one of the most standard and commonplace examples of an introspective exercise. However, limiting one’s scope to one core case for the sake of healthy conservatism has its own dangers: we lose sight of the larger picture, consisting in some of the less obvious introspective powers we have. Ultimately, if we are to gain a more complete understanding of introspection, we will have to consider some of the less standard cases of it. And it is this more complete understanding that I aim at here.

And second, as we widen our focus to cases beyond the core ones, our intuitions about what counts as introspection will start to fail us. With that in mind, we need to reconsider how narrowly to define introspection at the outset, such that non-core cases are included. In particular, for those of us who think it definitional of introspection that it is special, or privileged in some sense, this may not turn out to be true of introspective cases outside of the core. As such, instead of starting with a very specific conception of introspection and ignoring all other related cases, I plan to let the defining properties of introspection emerge as we analyze the wide variety of relevant cases. So rather than letting conceptual analysis dictate which cases are relevant, what introspection is exactly and whether it is in any sense special or privileged is better left for the time when we have a more complete sense of the various introspective feats we can accomplish.

**Terminological Issues**

When it comes to the topic of introspection and self-knowledge, there are no firm conventions with regard to the use of even the key terms. In particular, the terms
introspection and self-knowledge themselves are used in different ways. The only common ground in the literature seems to be that introspection offers one path to self-knowledge, but beyond that, there is little agreement.

Nonetheless, there are two common ways of delineating the body of self-knowledge: by its content, and by the way in which its content is acquired. Taking self-knowledge to be simply the body of knowledge about oneself is the broader definition. The narrower definition restricts self-knowledge to the body of information one can only acquire introspectively, i.e. in a uniquely first-personal way. Clearly, the narrower definition is highly objectionable as an all-things-considered take on self-knowledge, since it is quite common for some of the introspectively acquired information to be also non-introspectively available. For example, one can discover herself to be in a good mood simply through feeling so, but also by catching oneself smiling for no apparent reason. How this information was acquired – by noticing one’s feelings or one’s smile – does not seem like the sort of factor that should influence whether the discovery of one’s good mood counts as being part of one’s current body of self-knowledge or not.

However, the broad definition of self-knowledge as the body of information about oneself is also problematic. Although it fits the literal meaning of the term, the interest in the topic of self-knowledge is generated by its largely mysterious status as, intuitively, epistemically superior to other kinds of knowledge. However, there is nothing mysterious or epistemically superior about some facts and how we acquired them, such as our birth date, place of birth, height, eye color, etc. And yet, these facts are included in the body of self-knowledge on the broad definition.
In response, one might attempt to tweak this broad definition and create a third and different way of delineating the relevant body of self-knowledge – as the knowledge of our mental attributes, rather than our physical ones. However, this definition is also objectionable, since the knowledge of your body’s configuration acquired via proprioception can be just as mysterious and epistemically superior (in the relevant sense) as knowledge of our own beliefs.

It is therefore difficult to provide the exact necessary and sufficient conditions for what counts as self-knowledge, and consequently what is introspection, since the two are so closely related. Nonetheless, it would be difficult to proceed without some sense, however minimal, of the kinds of phenomena I am attempting to account for. To that end, I shall make a couple of distinctions. Throughout, I use the term \textit{self-knowledge} in its literal, broad sense, signifying the body of information about ourselves. Given its breadth, this concept will be of very little use here. A more interesting concept for my purposes is that of \textit{introspective self-knowledge}, which is the body of information about oneself (actually) acquired through introspection, however “introspection” is defined. But the most useful concept for my purposes is that of \textit{introspectively knowable (or available) facts}, which is the body of information that is at least possible to acquire through introspection (however defined). It is this last concept that I think most closely captures the set of phenomena that are of most interest to philosophers, and as such, it provides the anchor for all discussion to follow. In particular, my goal is to determine how such facts become introspectively available, and from there I construct an account of introspection that is, ultimately, pluralistic.
What makes the relevant facts available *introspectively*, as opposed to non-introspectively, is particularly hard to say. The difficulty in defining *introspection* is in large part due to the fact that the body of introspectively available facts seems to be extremely heterogeneous. It is simply extremely unlikely that the wide range of facts that make up the kinds of things we can find out about ourselves through introspection can all be accessed in exactly the same way, and with the same degree of epistemic security and accuracy (a point which I explore further in the section immediately to follow). As a result, we may have a good sense of the core cases of introspection, but these are far from sufficient to give a complete definition of what introspection is, since the body of introspectively knowable facts is much wider than the core.

This difficulty is quite apparent in the classic definitions of *introspection*, too, which are only approximations of the relevant phenomenon. For example, David Armstrong (1968) describes introspection simply as “awareness of our own state of mind” (95); and similarly, William James – to him, introspection is “the looking into our own minds and reporting what we there discover” (1890, 95). More often than not, this already limited and perceptual-sounding definition is restricted even further, by pointing to some of the alleged features of this awareness, such as its uniquely first-personal, or private, status, and its (frequently, resulting) epistemic privilege (e.g. Descartes 1641).

However, such restrictions are only helpful if we are investigating what, if anything, is special about introspection as a route to self-knowledge – i.e. in a purely epistemic endeavor. By contrast, my primary goal is not epistemic, but metaphysical – to gain a more complete understanding of what introspection is, and how it functions. Whether it is in any sense special, epistemically or otherwise, is something that will become apparent
from the discussion of the nature of introspection. And it is far from clear that the usual restrictions, such as privacy and epistemic privilege, apply to introspection across the board, outside of the core cases. For example, one might consider their own mental life in the same way as a therapist might do for someone else, with the same degree of objectivity. Being one’s own therapist is not a core case of introspection, but it is hard to deny that such an exercise is at least partially introspective. And yet, taking on the role of one’s own therapist enjoys none of the privacy or privilege that a core case might.

In fact, when considering the outlying cases of introspection, the line between introspective and non-introspective access to one’s own mental life becomes noticeably blurry. Even testimony – usually taken to be the paradigm of non-introspective access – can contribute to a case that overall seems introspective. For example, after several sessions a therapist might tell a patient that she is still grieving the loss of her childhood pet, and the patient might believe that and attempt to internalize that assessment. However, the patient finds accepting this assessment difficult, and thereby discovers that the therapist is in fact mistaken – the patient has actually dealt with that loss a long time ago. In this case, the therapist’s (mistaken) testimony is a key component of the patient’s introspective exercise, without which the relevant discovery would not be possible. Of course, this is not meant to conclusively show that acquiring self-knowledge through testimony can be an introspective exercise, but rather illustrate the difficulty in drawing the line between introspective and non-introspective forms of access, outside of the core cases.

Similarly, memory might be commonly excluded from the discussion of introspection, but on further reflection, it is not obvious that it is not an introspective
source of self-knowledge. Although I do no discuss memory as a potential source of introspective self-knowledge in any detail, it is worth noting that there is no good, principled reason to exclude it from the list of introspective routes to one’s mental states. One might be tempted to appeal to the fact that recalling something brings to mind only one’s past experiences and mental states, so it grants no access to one’s current mental states. However, the definition of introspection rarely includes a restriction on access to one’s current states only – and for good reason, since most of the mental states we are happy to call “current” stretch out well into the past, like many of our standing beliefs. Moreover, the feeling of familiarity associated with memory appears to be itself introspectable, and sometimes shared with other clearly introspectable experiences, such as the experience of déjà vu. Although I do not take a stand on whether memory should or should not be included when we are thinking of how to define introspection, what I take these considerations to show is that the boundaries of introspective access to mental states ought to be controversial.

For these reasons, and in order to gain as complete an understanding of introspection as possible, I find it more productive to reduce the number of definitional restrictions on introspection. Throughout, I will assume introspection to be simply a first-personal way of accessing one’s mental states, but not necessarily unique, or privileged in any sense. Whether it is in any sense privileged will become apparent as we examine cases that are intuitively relevant, developing a clearer view of introspection and its features from there. Let us now turn to the wide variety of introspectively knowable facts.

Starting with Examples
As already mentioned, there are many things we can know about ourselves starting from facts about our place of birth, our height, eye and hair color, to more complex and arguably less publically accessible facts like our past experiences, our plans for the future, our hopes and fears, etc. One of the few restrictions on this discussion of self-knowledge is that our knowledge of the kinds of facts mentioned first – place of birth, names of your parents, your height and skin color – is going to be basically ignored. These are not introspectively knowable facts in that we can learn them only in the very same way as other people – we are either told by, e.g., our parents or read the relevant documents (as in the case of our date and place of birth), or we observe them in the mirror (as in the case of our physical attributes). Importantly, one must learn about one’s place of birth via testimony or other indirect evidence (like seeing your birth certificate), and similarly one must learn about one’s eye or hair or skin color by seeing yourself in a mirror (or again via testimony, as in the case of blind people). One cannot acquire such self-knowledge without resorting to publically available evidence. The same cannot be said of, e.g., one’s plans for the future, one’s desires, and particularly of one’s beliefs. These are the facts that are, intuitively, introspective knowable.

On one end of the spectrum of introspectively knowable facts, we find the core cases of access to our conscious (and occurrent) mental states. Our current sensations and perceptions are good examples of such states – the headache I am feeling right now; the rose bush I am seeing outside the window; the “pins and needles” sensation in my left foot; the itch on my back, etc. I know I have all these experiences and I do so apparently effortlessly, and without needing to consult anyone. There are also other mental states I may be currently having that I know about. For example, I believe I am seeing a rose
bush and not a very colorfully dressed person outside; I also now have the desire to be sitting by that rose bush rather than stay indoors; I am also hoping that this may soon be the case, but I am doubting that the weather will be permitting later on, and so I intend to take a break to enjoy the weather and the rose bush before it starts raining, though I would prefer to have had more work done before going outside, etc.

This collection of attitudes towards propositions involving the rose bush is only the most salient of my mental states (in that they are at the forefront of my consciousness), but there are other mental states I can become aware of given appropriate prompting, some less fleeting than others. In particular, I might become aware of my standing belief that I can speak Russian and that I can count by, say, being prompted to answer a question about the number of roses I see, in that language. Or I may become re-acquainted of my dislike for overly sweet foods by biting into a slice of flourless chocolate cake. I may also know that I am in a chatty mood, should I discover that I’m itching to tell anecdotes to whoever will listen. In fact, I could be apprehensive about what I might say, feeling so chatty. And should I inadvertently say something not entirely appropriate, I’d be feeling disappointed in myself for not being able to filter my conversation better.

Along with these standing or at least longer lasting mental states, there are also our baseline dispositions (sometimes termed “affect”) to react agreeably or disagreeably to various events in our lives, which we may know about. For example, an excitable person might become very happy should she get some good news, while someone more gloomily disposed might look for caveats after hearing the same good news. It is certainly harder to know what one’s baseline dispositions are, but it is also a kind of self-knowledge one
may (and usually does) eventually acquire, sometimes introspectively. Similarly with character traits – one may learn that one is kind, sarcastic, punctual, cheeky, generous, funny, etc. Offhand, it is not as easy to learn that one is, say, courageous than to learn that one sees a rose bush outside, but one can nonetheless have both kinds of self-knowledge and everything in between.

Besides having access to such a wide variety of mental states and their contents, we can also access other, less obvious or discussed properties of our mental life. Being able to tell a belief from a desire from a sensation, and what content, if any, that mental state has – which are the abilities the introspection literature focuses on – make up only a fraction of the kind of information about our mental states we can access. We can also introspectively find out about the duration and continuity of our mental states (how long we’ve had it for and whether we ever stopped having it, which is particularly amazing for sensations); the intensity or the level of commitment to a mental state (e.g. how much my head hurts, or how much I want ice-cream, or how much I really believe that my friend is going to show up on time); our comfort level with having a particular mental state (I’m thinking of the associated, largely generic phenomenology of, for example, being annoyed at wanting more cake, or being proud of your belief that you did a lot of work today, or being surprised at your enjoyment of a country song); the level of normativity, or compulsion to result in behavior a mental state has for you (e.g. that you really ought to act on your belief that god doesn’t exist, but that it’s not so important that you act on your desire to go on the roller-coaster ride); etc. – the list goes on and on.

The bits of self-knowledge we may acquire mentioned above are just a small sample of all the things we learn (and retain) about ourselves throughout our lives. That much is
relatively uncontroversial. However, exactly how we acquire all this self-knowledge is far from clear. But focusing for now on the richness and the variety of our mental life, it should strike us as absolutely amazing that we can know so much about it, granting even that much of it remains inaccessible to us, and that we are able to assemble a relatively coherent and, even more surprisingly, robust self-image. This bird’s eye view of the variety and richness of the kind of information about ourselves we can apparently access is the starting point and the heart of the argument for pluralism.

**Pluralism about Introspection**

The goal of this dissertation is simple – to argue for pluralism about introspection. But before outlining my arguments, it is worth explaining what pluralism is in a bit more detail.

Pluralism about introspection can be treated both as a meta-view and a particular first-order view of introspection. It is a meta-view in the sense that, according to it, introspection can be accomplished in more than one way, each of which receiving a distinct analysis. Stated this broadly, pluralism does not make any claim with regard to the specific analyses, or even their number (as long as it is greater than one), that introspection can receive. And perhaps this meta-pluralism is not very controversial, as long as one is willing to be flexible with how one defines *introspection*.

By contrast, a first-order pluralist account does involve commitments to the particular analyses of introspection, their number, and the range of mental states to which each analysis applies. First-order pluralism is the more interesting and controversial type of pluralism, compared with meta-pluralism. There is a fair amount of variety in the
possible first-order pluralist accounts one might want to defend. The most obvious way in which first-order pluralist accounts can differ is in the particular analyses of introspection they offer. For example, one pluralist account could hold that propositional attitudes, and beliefs in particular, are accessed through a reasoning-based mechanism, and emotions are accessed “perceptually”/”observationally” – while a different pluralist account might claim that propositional attitudes are non-mechanistically accessed but rather constitutively available to us in virtue of us being rational agents, though our access to emotions might require a reasoning-based mechanism.

The metaphysical explanation of our access is only one way in which first-order pluralist accounts may differ. Another interesting and important way in which such accounts can diverge is in the range of mental states each of the proposed analyses is meant to explain. For example, two pluralists can agree that introspection is accomplished both through reasoning and perceptually, but they could disagree on which mental states are accessed through reasoning and which are accessed perceptually. As in the previous example, one pluralist account could include reasoning-based access to beliefs and perceptual access to emotions, while a different pluralist account could include reasoning-based access to all propositional attitudes, not just beliefs, and perceptual access to sensations, but not emotions. A particularly interesting type of pluralism, to which I am partial, is one according to which different analyses of introspection can have overlapping ranges of application. For example, a pluralist could hold that all propositional attitudes could be accessed through reasoning, while motivational and phenomenology-rich states, such as desires and emotions, are accessed
perceptually; in this example, desires can be accessed both through reasoning (in virtue of being a propositional attitude) and perceptually (in virtue of being a motivational state).

These two factors, the proposed analyses of particular introspective mechanisms and the range of mental states to which they offer access, determine the available first-order pluralist accounts. In particular, these factors determine the number of introspective mechanisms a pluralist account includes, and as such, ultimately shape the extent to which introspection receives a unified account. Although a pluralist account does not naturally fit with a view of introspection in which it is a homogenous or unified ability, it is nonetheless possible to have a sufficiently minimal pluralist view which is compatible with a largely unified view of introspection – and would thereby allow for sweeping claims concerning the potential privacy and epistemic privilege of introspection as a whole. However, since the unity of introspection is so highly dependent on which first-order pluralist account one wishes to defend, I leave the discussion of which properties introspection might have as a whole until the particulars of my preferred pluralism account are settled.

**Reasoning vs. Perception**

Turning our attention now to the elements of a pluralist account of introspection, there is a distinction to be made between the available accounts of introspection, which plays a significant role in this dissertation. Among the first-order views offered in the literature on the nature of introspection, there is a broad divide between those that model introspection on reasoning and those that model introspection on perception, which I term *reasoning-based* and *perceptual* models respectively. Setting aside the occasional
Constitutivist account of self-knowledge, like Shoemaker’s (to be discussed below), on which our beliefs about our mental states, and beliefs in particular, simply are one and the same as those (first-order) beliefs, the divide between reasoning and perceptual models of introspection largely exhausts the existing views on the matter. Much hangs on which side of the debate one chooses to defend, but it is not always clear whether there is a distinct set of differences between the two sides; at least, different sorts of considerations matter for different philosophers. And since I plan on bridging some of the gaps between the two sides, it is worth getting clear on the differences between the two sets of views.

There are four differences between reasoning and perceptual models of introspection that are particularly salient, which I will rely on throughout my arguments. The list I am about to present is neither exhaustive, nor does it provide the necessary conditions for belonging to one side of the debate and not the other. Rather, the list is a guide to (in some sense) typical features and desiderata on both sides; depending on one’s views on related issues, like the nature of inference, consciousness, mental contents, justification, and cognitive mechanisms, one’s view on introspection may not exhibit all four features of, say, a reasoning-based model of introspection – in fact, it may borrow a feature or two from the perceptual side. However, working with the default (or perhaps naïve, or even toy) views on each of those related topics will allow us to draw a relatively clear line between the two sides of the debate on the nature of introspection, which will be helpful throughout my arguments. With these caveats in mind, let us now examine the four differences between reasoning and perceptual models of introspection.

(i) The need for a dedicated cognitive mechanism
This is the most significant difference between reasoning and perceptual models, which is that perceptual models tend to require a dedicated cognitive mechanism, while the reasoning models do not. What counts as a dedicated cognitive mechanism is easy enough to gloss – it is some amount of cognitive “hardware,” the sole purpose of which is accessing or tracking (a subset of) one’s mental states, which would otherwise be absent. The most vivid examples of such a mechanism are the “inner eye” metaphor taken seriously, as in David Armstrong’s “brain scanner” commitment. By contrast, an account of introspection that does not require a dedicated mechanism can explain our access to our mental states in reference to cognitive mechanisms that already have some other function, such as general-purpose reasoning, which a normal adult agent is expected to have regardless of her ability to introspect. What this distinction amounts to more precisely, though, is far from clear. In particular, some reasoning accounts, such as Byrne’s (2005), require that we follow a particular rule of reasoning, which we presumably store, in some sense. And even though this rule is only useful for introspection, it does not seem as though Byrne (and defenders of similar views, like Gallois 1996) count that rule as a dedicated cognitive mechanism. Perhaps the distinction here is one of degree – that is, of the complexity of the required cognitive mechanism – but even so, it is far from clear that the extremely simple “monitoring mechanisms” proposed by Nichols and Stich (2003) are any more complex than Byrne’s rule-following account. In any case, there is certainly a sense in which perceptual models tend towards requiring a greater amount of dedicated “hardware” than the reasoning models; though it is worth flagging that particular perceptual models may not be any more cognitively
demanding than particular reasoning models. This theme will be recurring throughout my arguments.

(ii) Types of inputs and outputs

Given that reasoning, and drawing an inference in particular, is usually conceived of as being a particular type of transition between two beliefs, reasoning models of introspection tend to require contentful states as inputs. Charitably conceived, reasoning models can account for access to states other than just belief, though I will argue that this is false. Similarly charitably interpreted, the outputs of reasoning models of introspection are not merely the contents of introspected states, but also the state types (e.g. belief, desire, hope, etc.); but again, I am not convinced that information beyond content is accessible through reasoning. In any case, there are no similarly principled restrictions on perceptual models in terms of their inputs or outputs. In particular, a given state need not be contentful, and that content need not be conceptualized, in order for all of it to be accessible via a perceptual mechanism. Of course, specific perceptual mechanisms may be extremely limited in their inputs and outputs – as can particular reasoning models, for that matter – so the relevant difference between the two types of models exists solely in virtue of the fact that reasoning is usually thought of as a very specific kind of (belief-limited) activity. Depending on how flexible one’s concept of reasoning is, the difference in inputs and outputs between the two types of models may shrink significantly.

(iii) Conscious, person-level accessibility

Since reasoning is usually thought of as the sort of thing that can be done consciously and on the person-level, reasoning models of introspection tend to allow for that
possibility as well (though they do not require it). By contrast, nothing of the sort is guaranteed on perceptual models; instead, the process of introspection is completed largely on a sub-personal level, and only the outputs can be conscious and person-level. There are two important points to flag about this distinction. One is that it does not concern the automaticity of introspection; that is, the claim that perceptual models work sub-personally does not entail that they are *constantly* working – they can (and often do) require a trigger, just as the reasoning models require. Similarly, the distinction does not concern the degree of “effortfulness” introspection may or may not have. Whether a process is or can be conscious or person-level – or otherwise – does not entail a thing about the amount of cognitive effort is required. The distinction I am drawing in this section is purely in terms of *format* that introspection can take. And in particular, the claim is that reasoning models describe introspection in such a format that allow (but do not require) the entire process to be person-level and consciously monitored, while the perceptual models tend to require some occlusion of the process from the subject’s consciousness, which allows only the outputs to be person-level and consciously accessed.

(iv) Epistemic responsibility

The final distinction stems, once again, from how we intuitively think about reasoning – as a process that is rationally assessable, and one that can influence the epistemic status of its output. In particular, one can be blamed for forming a false belief due to bad reasoning, and praised for a forming a true belief due to good reasoning. So it is with reasoning models of introspection – the results of introspection they give, the particular beliefs we form in the process, are rationally assessable and epistemically
significant, simply in virtue of the process that led to it. This does not seem to be true of perceptual models though; at least they do not carry an equal burden of epistemic responsibility. If the dedicated introspective mechanism malfunctions on a sub-personal level, intuitively one has done no epistemic or rationally significant wrong, or at least to the same extent as if one reasons poorly. Importantly, this claim is made purely on the level of intuition, since one can favor an account of justification – process reliabilism – that places an equal epistemic burden on a malfunctioning dedicated mechanism as on bad reasoning. However, having to explain away the intuition I just described is also the burden of a process reliabilist, and so I leave this distinction at the level of intuition.

With these four differences between the reasoning and perceptual models now on the table, it should be clear that their force is largely cumulative. Given that each difference is a matter of degree, and sometimes depends largely on one’s perspective on the relevant issue, there is no single, definitely way to draw the line between the two types of models – but this should not be surprising, since it is equally difficult to call on one clear and simple difference between reasoning and perception, in general. Since the line between reasoning and perceptual models is admittedly fluid, it does not seem possible to argue in favor or against either side en masse, as providing a successful or an inadequate account of our access to a given mental state type. However, when a specific representative of a reasoning or a perceptual side succeeds or fails to account for our access to a given mental state type, I shall supplement my explanation of that fact with the above-mentioned distinctions, when helpful. Thus the differences between the reasoning and the perceptual models will not play a crucial role in my arguments, thought they may aid in
understanding why we ultimately need both sides to provide a complete account of introspection, or so I argue.

**The Arguments for Pluralism and Dissertation Outline**

In this dissertation, I defend a particular first-order pluralist account of introspection (which, of course, also entails the broader, meta-pluralist claim). Although completeness of the account of introspection is my ultimate goal, I do not offer a pluralist account that is in fact complete here, simply in virtue of how vastly incomplete our understanding our mental life currently is. So although I take into account the wide of range of mental states that we can access, for now I only claim to explain how we access some of them. Still, the range of the mental states I consider is greater than the range of mental state to which accounts of introspection usually apply – and even with this incomplete picture of our mental life, pluralism is necessary. To that end, I offer two main arguments.

The first argument involves examining all the main existing models of introspection, and assessing the extent to which they can generalize to provide a complete account of introspection. There are four main views on this topic I will discuss here, which I will label “Constitutivism,” the “Third Person Model,” the “Transparency Model,” and the “Perceptual Model”\(^1\), dedicating a chapter to each of the views, except Constitutivism (which is dealt with in the next section). These four views represent the most common,\(^1\) The labels “Constitutivism” and “Perceptual Model” (aka “Observational View”) are pre-existing, though not universally used, while the other two – “Transparency Model” and “Third Person Model” – have been suggested to me by Alex Byrne as more neutral names for the relevant views than those which are more frequently used to designate them (“Transparency Model” is the label I’ll use for the view also known as the “Inferential Model,” or the “Ascent Routine” view, while “Third Person Model” is the label I’ll use for views like Gilbert Ryle’s and “Theory Theory”).

19
or at least the most discussed, ways of explaining the nature of introspection. There are also several views that focus on some of the alleged features of introspection, Richard Moran’s Commitment view being the most significant example; but these views do not present accounts of the nature of introspection, i.e. its metaphysical structure, and are therefore are of no immediate use. These views, and Moran’s in particular, are nonetheless interesting in highlighting the epistemologically significant features of introspection, which I will be in a position to discuss later, after its metaphysical commitments have been clarified. But right now I examine the existing (metaphysical) accounts of introspection so that a new and improved account can emerge. The goal is to pin-point the places where the existing accounts of introspection break down, and show that by carefully synthesizing the existing accounts together we can generate a different account of introspection that explains better than any of the existing accounts do individually how we can access as much of our mental life as we do.

Thus, chapter 2 is dedicated to the Third Person Model, chapter 3 to the Transparency Model, and chapter 4 to the Perceptual Model. In particular, I argue in each of these chapters that each of these models is necessary but not individually sufficient to provide a complete account of introspection. Constitutivism is discussed at the end of this chapter, which I argue to be unnecessary for a complete account of introspection.

After this three-chapter argument for my preferred pluralist account, chapter 5 is dedicated to a different, self-contained argument for pluralism, one that is based in a particular example. The situation I describe in that chapter is meant to be relatively common in everyday life and generalizable, though not as simple as the kinds of introspective examples that are more commonly considered. In particular, the
introspective example I describe involves several steps, and I argue that the differences between the steps are best explained through different analyses of introspection. Thus this case supports pluralism about introspection.

In concluding this dissertation, I consider some of the epistemic features of introspection that result from the defended pluralist account. Taken as a whole, introspection turns out to be neither unified nor clearly special in the relevant epistemic sense, unless one thoroughly restricts the scope of the relevant introspectively accessed facts.

With this plan in mind, let us now turn to the first argument for pluralism, and begin examining the existing accounts of introspection. The first view to consider is Constitutivism, and since I am ruling it out, rather than adding to my pluralist account, I dedicate only a section to it.

**Constitutivism**

Constitutivism is a view almost exclusively associated with Sydney Shoemaker, who originated it. It is a unique account in that it bestows an infallible way for rational agents with the right conceptual arsenal to acquire an impressive amount of self-knowledge at will, largely of one’s beliefs and desires. The entire account hinges on a single claim, that the relation between first-order beliefs (e.g. “I believe it is raining”) and the corresponding second-order beliefs (e.g. “I believe that I believe it is raining”) is non-contingent, according to Shoemaker. His defense of this claim consists in a *reductio* of the possibility of a condition he calls “self-blindness”. When an otherwise fully rational agent suffers from this condition, she is unable to verify her second-order beliefs in a
first-personal way, though her ability to do so in a third-personal way is intact, as is her
general ability to form beliefs. Shoemaker’s thought is that self-blindness is a conceptual
possibility only if the relation between first- and second-order beliefs is contingent. So
by showing self-blindness to be conceptually impossible, he hopes to conclude the
relation between first- and second-order beliefs is non-contingent:

To deny the possibility of self-blindness is to hold that it is implicit in the nature of
certain mental states that any subject of such states that has the capacity to conceive
of itself as having them will be aware of having them when it does, or at least will
become aware of this under certain conditions (e.g., if it reflects on the matter).
(Shoemaker 1988)

The focus of the *reductio* is on showing that a rational person whose only defect is the
inability to verify her second-order beliefs in a first-personal way would be prone to
irrational behavior. An example of such a behavior is failing to act on one’s beliefs and
desires, the least of which is failing to assert in the appropriate circumstances that one
believes that p (when one does in fact believe that p). But according to Shoemaker, a
rational person would not act in such ways. Therefore, either such a person is fully
rational and suffers no defect, or she is not fully rational. In either case, the original
hypothesis is contradicted, and therefore the relation between first- and second-order
beliefs can be assumed to be non-contingent.

This is the only argument Shoemaker has ever given for his view since its inception,
and though its conclusion has always been that first- and second-order beliefs are non-
contingently related, he has recently (2009) clarified the precise nature of this non-
contingency – it is the relation of partial constitution (hence the name of the view). That is, after enough restrictions have been specified, second-order beliefs are partially constituted by the relevant first-order beliefs, and it is this relation that allows us to enjoy the only first-personal access to our own (first-order) beliefs that we have. And as far as I can tell, there is no additional argument for the constitutive claim besides ruling out Peacocke’s view, which Shoemaker thinks others might consider a competitor to his view, though he himself does not seem to think so.

Constitutivism is a bold and elegant view, but I think it is ultimately unmotivated. One problem I will highlight with its sole existing defense – others have noted different problems\(^2\) – is that the desired claim, that first- and second-order beliefs are related non-contingently, does not seem to follow from the *reductio*. Even if we grant that self-blindness is a conceptual impossibility, which is the proper conclusion of the *reductio*, all this shows is that, necessarily, rational agents can verify their second-order beliefs in a first-personal way. But whether this ability is due to a non-contingent relation between first- and second-order beliefs, or to any other property of beliefs or anything else for that matter, is not established by the *reductio*. There needs to be a further argument justifying the step from it being necessary that rational agents can verify their first-order beliefs in a first-personal way (i.e. them not suffering from “self-blindness”) to the claim that such verification is due to a necessary connection between first- and second-order beliefs. As it stands, this step is unjustified. And I see no independent motivation for the claim that first- and second-order beliefs are non-contingently related.

There are, perhaps, other ways to argue towards the constitutive view, but I agree with the existing assessments\(^3\) that, whatever the argument supporting Constitutivism, the view borders on trivializing self-knowledge acquisition. The worry is that Shoemaker’s view relies on an overly high level of rationality – one that would qualify a person as irrational merely for once failing to become aware of one of her beliefs in the appropriate circumstances. And even if we set aside the empirical question of whether we are in fact such highly rational agents, which seems in any case unlikely, the question of how such highly rational agents acquire self-knowledge is much less interesting than the question of how the less ideally rational agents do it.

Another problem is with Shoemaker’s positive view about belief structure, which has extremely limited use. In particular, I can see no benefit from the view that first-order beliefs partially constitute second-order beliefs beyond granting us infallible access to certain beliefs (and perhaps desires, though how it works for desires is never specified). And in general, it does seem appropriate to demand some independent motivation for a revisionary view of mental structures, like beliefs, even if this motivation is merely phenomenological – which, of course, is unavailable in this case since Shoemaker himself does not think that beliefs have distinct phenomenology. This is, perhaps, the root of why there is no obvious alternative argument for Shoemaker’s view I can see besides the one he gives from “self-blindness.”

As a result of these considerations, I have insufficient independent motivation to defend Constitutivism and so to include it in a pluralist account of introspection. Instead

\(^3\) E.g. Gertler (2003), Rey (forthcoming).
I now turn to the models of introspection that do not rely on any special views about rationality or belief structure. In the next chapter, I turn to the Third Person Model, and argue that it is necessary element of a complete view of introspection, but it is not by itself sufficient to explain all of our introspective access.
Chapter 2: The Third Person Model of Introspection

So far, I have argued again Constitutivism as a necessary part of a complete account of introspection. There are now three models remaining to be examined in a similar manner. With all three models, I argue that each is a necessary but not individually sufficient to provide a complete account of introspection across all mental state types, thus necessitating pluralism about introspection. Of the three remaining models, I will first consider the two models that do not involve any cognitive mechanisms dedicated to introspection, and instead rely on cognitive mechanisms that have some independent function. Following Byrne (2005), I will call such models “economical.” The two economical models are the Third Person Model, according to which introspection is accomplished via the same mechanism we use to determine other people’s mental states; and the Transparency Model, according to which introspection requires nothing more than general-purpose reasoning, used in a strictly first-personal way.

In this chapter, I focus on the Third Person Model – an umbrella view that ties introspection to observation of behavior to some degree, the most notable incarnations of which are the Rylean view of introspection, and, more recently, the “theory theory” account of mental state attribution. There are two main goals for this chapter. The primary goal is to show that a view like that accurately but only partially accounts for our access to our mental states, and is therefore necessary but not sufficient to explain our introspective access to all mental states. The other goal is to argue that a view like that is worth thinking of as an account of introspection, as opposed to non-introspective route to self-knowledge. Taking these goals in turn, I first describe the Third Person Model; next I
defend it against several objections; and finally, I argue that it is worth thinking of this model as a genuine model of introspection.

**The Third Person Model**

The Third Person Model is the family name I’ll use for views, the core commitment of which is that there is no contrast between 1st- and 3rd-personal acquisition of beliefs about mental states. In other words, however we acquire beliefs about other people’s mental states is also how we acquire beliefs about our own mental states. To get a better sense of this and other commitments of the Third Person Model, let us look at its most representative member – “theory theory,” a view commonly associated with Alison Gopnik.

“Theory theory”, which is sometimes claimed to be the leading scientific view of mental state ascription, is usually thought of exclusively as a view of belief acquisition concerning other people’s mental states, not our own. That is, “theory theory” offers an account of how we are able to acquire beliefs to the effect that, for example, my neighbor believes it is raining, rather than my believing that to be the case. And though this is an important contrast, “theory theory” is inevitably, even if secondarily, a view about introspection as well, since whatever cognitive mechanism accomplishes 3rd-personal acquisition of mental state beliefs has to be the mechanism for introspection too on “theory theory,” in light of its core commitment.\(^5\)

---

\(^4\) E.g. Stich and Nichols (2003).

\(^5\) Whether “theory theory” is, in any sense, an account of introspection is something I discuss later in the chapter.
Turning now to that cognitive mechanism, there is some disagreement between defenders of “theory theory” over its the exact nature, but its basic structure is the same on all versions of “theory theory”. The way in which we acquire knowledge of other and our own minds on the “theory theory” approach is by making use of the stock of psychological rules that link environmental inputs, mental states, and behavioral outputs. These rules make up the “theory of mind” (ToM) mechanism, or module. And by relying on these rules we can triangulate the current mental states of others, as well as ourselves. This triangulation is accomplished by reasoning using the ToM rules. For example, if you want to find out what your friend believes about the weather, you would observe her behavior, part of which might include carrying an umbrella though not currently using it, and also note the environmental conditions, which may involve the absence of rain – and after that you would infer that your friend believes that it is not raining but it might later today. So on “theory theory”, introspection turns out to be (a) a reasoning-based process that (b) relies on information about the links between the environment, behavior, and mental states, where neither (a) nor (b) are unique or dedicated to introspection.

6 As with any broad view like “theory theory” there are significantly different versions of it – in the case of “theory theory” there are internal and external versions. The external version deals with how intentional terms like belief and desire function, while the internal version specifies a cognitive mechanism that underpins our ability to attribute mental states to ourselves and others. In fact, both versions are compatible with each other, but only the internal version of “theory theory” is of any use in a discussion of introspection, so from now on I’ll reserve the term “theory theory” for its internal version.

7 I feel I owe an explanation of why I am not also taking up the “simulation” theory, since these two theories are usually set up as directly contrasting each other. The reason for omitting the simulation theory is that the simulation theory is just a framework for thinking about introspection, according to which introspection does indeed require a cognitive mechanism distinct from the one we use to find out about other people’s mental states. That is, all that simulation theory states is that there is a contrast between 1st- and 3rd-personal acquisition of beliefs about mental states, which “theory theory” denies. But
It is now important to contrast “theory theory” with the Third Person Model, which is an umbrella view with significantly fewer empirical commitments than “theory theory.” The main difference is that the Third Person Model does not involve any claims about the involvement of cognitive modules, the precise structure of various cognitive mechanisms, and other controversial issues in cognitive science. The Third Person Model is only committed to the undeniable fact that motivates “theory theory,” namely, that we are capable of ascribing mental states to someone just by observing their behavior, and that someone could be oneself. How exactly that is accomplished, which is the focus of “theory theory,” is not specified on the Third Person Model. It is surely reasonable to expect that this is accomplished with the help of our general-purpose reasoning ability, and perhaps it also involves a dedicated mental state ascription mechanism – though not dedicated to mental state self-ascription as such – but beyond these general claims the Third Person Model is non-committal. In light of such an abstract perspective on introspection that the Third Person Model offers, it should now be clear why Ryle’s view fits into this category, since his claims about introspection were largely negative, i.e. denying that introspection is any different from how we ascribe mental states to others.

There can be little doubt that we can and do determine mental states, both our own and other people’s, from behavior. I can learn that I am afraid of heights by observing that I keep refusing to take part in activities like skydiving and paragliding; or I can learn that I really like coconut more than I thought by consistently choosing dishes that contain it; or that I am in a good mood by catching myself smiling for no apparent reason and the simulation theory involves no claim about the specific mechanism that fits this description. And insofar as the simulation theory offers no specific model of introspection, it is of no use here.
saying particularly nice things, etc. The only problems that might arise for the Third Person Model concern the range of its applicability, and whether it can be considered the *only* route to self-knowledge, as some “theory theory” defenders claim. As I discuss in the next section, the Third Person Model cannot be our only introspective source of self-knowledge.

**Objections to the Third Person Model**

There are three sets of objections to the Third Person Model of introspection worth discussing, largely because they highlight some important and easily misunderstood features of the view. I present the objections in the order of increasing difficulty they present for the Third Person Model, even though I show that the worst of objections can only impose a limit on the applicability of the model, rather than expose it to be fundamentally flawed.

*First Objection: “Introspection does not involve observing behavior”*

The first set of objections stems from a naïve intuition about the proper use of the term “introspection,” namely, that if one self-ascribes mental states *by observing one’s own behavior*, one is not introspecting. This intuition can lead one to two conclusions. The first is that it is simply wrong to say that one is introspecting when one is trying to figure out whether she believes that it will rain by seeing whether she will take an umbrella when she goes out. That just is not something people do, or so the objection goes.

This line of reasoning presupposes that the Third Person Model, or any purported model of introspection for that matter, carries the burden of describing the *only* way of
self-ascripting mental states. If that were the case, then indeed the Third Person Model would need to explain and apply to every case we would be willing to call “introspection.” And perhaps that is a commitment that some “theory theorists” make. But the Third Person Model as I have described it involves no such commitment, and is therefore under no obligation to account for certain cases, such as knowing that one believes that it will rain. By contrast, the Third Person Model accounts very nicely for cases in which one finds out whether one is a good parent or son or daughter, or whether one believes in god, or whether one is happy living where one does, or whether one really likes one’s neighbor. This absence of a claim to completeness on the part of the Third Person Model is a notable departure from “theory theory,” some versions of which are specifically tailored to rule out the possibility of any other forms of introspection (a commitment that is, once again, most closely associated with Alison Gopnik, e.g. Gopnik and Wellman 1994, and Gopnik and Meltzoff 1997). The denial of completeness is a key feature of my version of the Third Person Model, as well as other models of introspection I examine, and as such it is worth defending. My reason for resisting the commitment to completeness, as can be seen from the first chapter, is that it is simply unjustifiable prior to examining the entire range of our introspective abilities, which is hardly a completed scientific project.

Returning now to the naïve intuition that observing one’s behavior is not part of introspection, the other conclusion it leads to is that even if we self-ascribe mental states in the way the Third Person Model describes, it cannot be said to be a model of introspection because introspection should not involve (external) perception, as it does
when we observe behavior, our own or someone else’s. Even the word *introspection* suggests that only “inner-directed” observation is relevant.

There are two things to say in response. First, this seems to be a good point at which we should treat the naïve intuition only as a loose guide to the core use of the word “introspection”. For it is far from clear that our hesitation to call the Third Person Model a model of introspection has any metaphysical consequences, rather than merely linguistic ones. Distinguishing between the merely linguistic practices and the metaphysically significant linguistic practices is perhaps not a unique but a particularly salient obstacle to the study of introspection, since, after all, there is hardly a way to observe or even confidently infer the presence of an introspective process, which we can then match with the appropriate linguistic practices. Ultimately, the word “introspection” seems to be as much in need of a guiding definition as it is a guide itself to its extension.

Second, even if we were to treat the naïve intuition as a guide to the metaphysics of introspection, all our hesitation to call the Third Person Model a model of introspection can really show is that the Third Person Model does not comfortably account for the cases we tend to think of as core. But that hardly establishes that the Third Person Model is not a model of introspection at all, unless we saddle the Third Person Model with the commitment to completeness, which I already rejected. Rather, this consideration might entitle us to treat the Third Person Model as a model of fringe cases of introspection – since it does account for some cases of mental state self-ascription – which is just as necessary when the point is to provide a *complete* account of introspection.
As one last and most efficient line of response to the thought that observing one’s behavior cannot be part of introspection it is worth pointing out that the Third Person Model, as I have described it, does not necessarily involve actual visual observation of one’s own behavior. The core of the model is the claim that mental states are inferred from behavior – but in the case of one’s own behavior and mental states, the behavior need not be (visually) observed as such. In other words, one’s behavior need not even actualize in order for something to be inferred from it. This possibility is entirely alien to “theory theory,” because it would require independent access to, say, one’s intentions to behave in a certain way, from which one can go on to infer something else about one’s mental states. And such independent access to one’s intentions immediately undermines the claim that the Third Person Model, and “theory theory” in particular, can be the sole path to self-knowledge. But since this claim to completeness is not one I am defending – and in fact, I am interested in showing its negation – the possibility of inferring something from unactualized behavior is perfectly acceptable in this context. Importantly, the possibility or even a need for an independent access to our intentions, and one’s intended unactualized behavior, does not yet show the Third Person Model to be false, or even redundant, because it remains useful in self-ascribing certain mental states, such as those that are deeply hidden (like suppressed/repressed beliefs and desires), and those that are largely dispositional, like character traits. In fact, a case of introspection that involves unactualized behavior is the focus of one of the later chapters.

Second Objection: “Third Person Model doesn’t capture what introspection feels like”

Moving on to other objections to the Third Person Model, the second set of objections targets the feel or “what it’s like” to introspect, i.e. its phenomenology. If it were true
that even just sometimes we employed the same cognitive mechanism to ascribe mental states to others as to ourselves, then one would expect it to feel about the same whenever that happened, whether we use it to find out about our own states or some else’s. But that seems hardly ever true. There is something that I can best describe as felt obviousness when it comes to self-ascribing mental states, which is largely absent when we consider other people’s mental states. The best examples of this felt obviousness include every time you consider your current mental states, like the thoughts and desires that currently occupy your attention. Whatever this felt obviousness is, it is certainly not proprietary to introspection, and perhaps not even best described as “phenomenology” at all, except that it fits the Nagelean definition of “what it’s like” to be or do something. And perhaps this felt obviousness is just a spontaneous first-person assessment of one’s epistemic status, having a high level of certainty. But whatever it is, it seems to accompany only self-ascriptions of mental states, rather than their ascription to others, and this difference should be enough to make it doubtful that we ever rely on the Third Person Model to self-ascribe mental states.

There is something important to take away from this objection, but first let me soften its blow. The objection rests partially on an exaggeration of just how obvious self-ascriptions of mental states seem. First of all, not every introspective effort results in a self-ascription of a mental state that feels equally obvious. It is much more obvious to me that I believe myself to have been born in the Ukraine, than that I am happy to be a resident of the United States. So at the very least, this felt obviousness, whatever it is, is a matter of degree, even when only self-ascriptions of mental states are concerned.
Moreover, we can indeed determine other people’s mental states with some degree of obviousness, too. In fact, if the other person is well known to you, it is sometimes more obvious what some of their mental states are than some of your own. For example, it is less obvious to me that my mother is happy to be a resident of the United States than that I am, but it is far more obvious to me that my father is the happiest of all three of us in this respect. What this shows is that there is some overlap in degree of obviousness accompanying self-ascriptions and ascriptions of mental states to others, though it could be true that on average self-ascriptions feel more obvious than ascriptions of mental states to other people. This difference could be explained by the difference in the speed with which we can ascribe a mental state to ourselves as opposed to someone else. And this, in turn, could be explained by the greater amount of information about ourselves we have stored, and which speeds up the inference from behavior to the appropriate mental state. The important point here is that these differences, be in obviousness, speed, and amount of information stored about the subject, are insufficient to establish that a different mechanism is always used when we ascribing mental states to others as opposed to ourselves. What these considerations show is that there need not be a fundamental difference between some of our mental state self-ascriptions and ascriptions of mental state to other people; and a difference in degree can be easily accommodated by the Third Person Model.

Nonetheless, it is doubtful that the level of obviousness that accompanies self-awareness of our current thoughts can ever be reached by any ascription of a mental state to another person. This is indeed a troublesome point for the Third Person Model, but in its vicinity is an even bigger problem, to which I now turn.
The bigger problem concerns the phenomenologically-laden and possibly non-conceptual features of mental states. Even if we set aside the more sophisticated cases like seeing red and blurry vision, something as simple as knowing that you feel a twinge, or an itch in your arm – before you do anything to eliminate these sensations, as we often do – presents an explanatory difficulty for the Third Person Model. No matter how broad we make our definition of “behavior,” barring trivialization of the word, there would not be enough appropriate input for the cognitive mechanism that could result in our awareness of a twinge. What these phenomenological considerations show is that the Third Person Model is limited in another respect by not being able to explain our access to the phenomenology of our mental states. As I will show, there is only one account of introspection that is not limited in this respect. But what is important to keep in mind is that this is just a limit of the Third Person Model, and not an indicator of its complete demise. This model still holds true, as already mentioned, for dispositional states and some beliefs and desires; it has simply been shown to be individually insufficient to explain our access to all introspectable mental states.

Third Objection: “Third Person Model doesn’t have the right resolution”

The final set of objections to the Third Person Model concerns the discrepancy between the actual “grain” or “resolution” of introspection and the “grain” or “resolution” that the Third Person Model can allow for. Consider a scenario, in which you have an 11am meeting tomorrow. Suppose also that you intend to show up at least 10 minutes early for that meeting. We form similar intentions all the time, and there seems to be no problem in our ability to be aware of the exact situation we are trying to bring about – in this case, being at least 10 minutes early for an 11am meeting. In fact,
you show up 18 minutes early for your meeting. The problem this scenario presents is that there are very few normal behaviors that can exhibit our intention with sufficient accuracy, and yet we have no trouble identifying the exact intention even in the absence of those behaviors. Specifically, we can easily tell whether our intention was to show up 10 minutes early as opposed to 15 minutes early, even though our actual behavior might compatible with both interpretations. So in some cases, the actual “grain” of introspection is finer than the “grain” that the Third Person Model can account for.

In fact, the “grain” problem is just the tip of the iceberg. The deeper problem is that there is just not enough actual behavior a person can exhibit to match the number of mental states we apparently have access to. The amount of information we seem to have about our mental life is just too great to suppose that we acquire it all from behavior, which is so much more restricted. The extent to which our self-awareness outruns the available behavior is best exemplified in our abundant awareness of occurent thoughts. Becoming aware of one’s current thoughts and experiences is frequently free of any behavior that might betray those thoughts or experiences. One only needs to consider the classic example of armchair philosophy, in which one can be well aware of one’s thoughts and yet do nothing but sit in an armchair. Imagining, for reductio purposes, that one would need to exhibit behaviorally one’s current thoughts in order to become aware of them is also instructive – consider always having to speak out loud or write down your

---

8 Perhaps we can never be aware of strictly current thoughts, as has been pointed out before (e.g. Hill 2009, Schwitzgebel 2010). Instead, we seem to be aware only of our most recent but nonetheless immediately past thoughts, as our awareness of those thoughts seems to occupy all of our current attention. This is an important observation, but it can be safely ignored here, since the difficulty for the Third Person Model here discussed is no less pronounced in light of this correction.
own thoughts before becoming aware of them (either from hearing or reading them). This might be a useful exercise, particularly to organize and streamline a proof, for example, but it is certainly not a necessary step with all of our thoughts. And the same point can be made with the help of Putnam’s (1963) “Super-Spartan” case. For one can well imagine a person going out of their way to limit their behavior, which, nonetheless, would have (almost) no impact on how much information about their mental states they can access.

But once again, all this shows is that the Third Person Model is limited in terms of both the grain of mental states we can access and the frequency with which we can apply it. None of these examples show that the Third Person Model is never applicable. So, for all the problems I have raised for this model, all I have shown is not that it is outright false but rather that it is limited to mental states (or events, more broadly) that do not have a strong phenomenology but are strongly tied to behavior, and are not very fine-grained.

**Is the Third Person Model a Theory of Introspection?**

I have already touched on the issue of whether the Third Person Model is genuinely a model of introspection, but only as an outright objection to the view. If, however, we accept that we do indeed acquire self-knowledge with the help of the Third Person Model, at least in some cases – as I have argued – the same question can be asked again, though from a different angle. Deciding how to label the Third Person Model – as a model of introspection, or a non-introspective path to self-knowledge – can help us get
clearer on the difference between our mere expectations of what introspection might be, and the absolutely necessary, core features of introspection.

One of the most elusive and controversial features of introspection is its alleged uniqueness, or “specialness,” as a route to self-knowledge. The basic idea is simple enough: there is some intuitive contrast between learning something about yourself through introspection, and doing so by, e.g., looking at a photograph of yourself. Introspection is usually thought of as something inherently first-personal, something from which no outside observer could benefit. However, how and where to draw the line between introspection and other routes to self-knowledge has remained far from clear, and a topic of considerable debate. But instead of defending a particular interpretation of what the specialness of introspection comes to, I prefer to stay neutral on the very possibility of introspection being special in any sense, since it interferes with the study of the nature of introspection, as I have argued in the previous chapter. Importantly, I do not want to argue, as Ryle did, among others, that there is nothing special about introspection. Rather, I want to show that taking a stance – any stance – on whether introspection is in any sense special at all, prior to identifying the particular mechanisms of introspection, is methodologically disadvantageous.

There are two disadvantages in particular worth highlighting – one is methodological, and the other is a structural. The methodological disadvantage of defining introspection as special or unique is that all other routes to self-knowledge become of less interest, if not ignored. And yet, it is not at all clear whether those other, non-introspective routes to self-knowledge – whatever they are – are not similarly significant contributors to our body of self-knowledge. As a result, we run the risk of building a skewed picture of how
we acquire self-knowledge, and consequently what, if anything, is special about it, in contrast with knowledge of the external world. Instead, it would be more instructive to consider our main sources of self-knowledge, whether they are in any sense special or not.

But more importantly, defining introspection as special, prior to giving a description of the particular mechanism(s) responsible for it, is nothing short of begging the question against certain theories, the Third Person Model being one of them. For one, as we have already seen, it is very difficult to agree on what would make introspection special; what makes it special to some is not recognized as special by others. And rightly so, for the motivating intuition that draws the line between introspective and non-introspective access is blurry, and as such offers no significant support to one interpretation of “specialness” of introspection over another. And without a clear and strong intuition\(^9\) behind the “specialness” of introspection, it can hardly be used as a reason to reject a theory of introspection that does not offer an explanation of it.

As one last observation in support of this line of reasoning, consider the status of “theory theory” and even the Rylean view as accounts of introspection. Neither of them have offered any distinguishing marks of introspection. If “specialness” were a defining feature of introspection, both of these views would be considered essentially eliminativist, i.e. denying that we are capable of introspection. However, such a thing was never suggested, even as a criticism of these accounts, even though in any other area of

---

\(^9\) In fact, I may be granting too much in calling the motivating idea behind the quest for “specialness” an intuition, for even its defenders are recognizing as its source the Cartesian approach to philosophy of mind (e.g. Shoemaker).
philosophy this would be exactly the accusation such views would warrant. But it seems almost a conceptual violation to consider a rational agent incapable of introspection, which is exactly what the Rylean view and “theory theory” should be taken to be implying, if we consider “specialness” to be a requirement on introspection. But instead, these accounts are seen as challenging only one of the (alleged) features of introspection, its “specialness,” and not the very possibility of us introspecting. What this shows is that there is already some acknowledged room in how we think of introspection. And rather than “specialness” being a necessary, defining feature of introspection, we should (and already do, as I have just argued) think of it as a desideratum, which is apt to be rejected – and importantly, rejected much sooner than rejecting the possibility of us introspecting altogether.

In light of these considerations, I find that it is best to think of the Third Person Model as an account of introspection. It may not account for the cases many of us will think of as core cases of introspection, such as the armchair cases present – and if that is the source of one’s resistance to the idea of assimilating the Third Person Model to accounts of introspection, the point is well taken. However, just because the Third Person Model does not account for the core case of introspection does not mean it does not account for any cases of introspection at all. And my goal is not to account merely for the core cases of introspection, but for all of them, which cannot be done without the Third Person Model.

Conclusion
As we now turn to the other economical theory of introspection, the Transparency Model, an important lesson to take away from examining the Third Person Model is that it is unlikely that a single model can successfully account for all of our introspectively accessed information. That is, we ought to be prepared for the possibility that no single model of introspection is individually sufficient. As negative as this point may seem, it does have a positive result, granting a significant amount of resilience to models of introspection in the face of their apparent failure to account for some introspective phenomenon. Should a model of introspection be unable to account for our introspective access to some mental state or event, its demise should not be considered imminent (which, historically, has been exactly the dialectic the introspection literature has fallen into). Instead, we should treat that model of introspection as merely restricted in some specific way. This observation, as we have already seen to an extent, provides the basis for resisting many objections, and so should be kept firmly in mind as we consider all other models of introspection, with the Transparency Model being our next focus.
Chapter 3: The Transparency Model of Introspection

In the previous chapter, I have argued that the Third Person Model correctly describes our self-attribution of dispositional mental states, like certain deeply rooted beliefs and desires, and character traits. The strength of that account is that it is “economical” – it does not require any dedicated introspective faculty or mechanism, which helps the pluralist thwart concerns about cognitive “overcrowding” with introspective mechanisms. However, the Third Person Model does not generalize to all mental states for various reasons, so in order to have a complete account of introspection, we need to supplement it with other models.

In this chapter, I examine another economical, reasoning-based model of introspection, the Transparency Model. I argue that unlike the Third Person Model, the Transparency Model can explain our access to all beliefs, but that it does not generalize to mental states other than belief. And having examined both economical models of introspection by the end of this chapter, we need to turn to the Perceptual Model in the next chapter, if we want to explain our access to all introspectable mental states. Thus, pluralism is the correct, complete account of introspection.

More specifically, the plan for this chapter is as follows: First, I motivate and outline the Transparency Model. Next, I move on to examining the objections to this model, defending it from some, and agreeing with others. I then examine the possibility of generalizing the Transparency Model to mental states other than belief, such as intentions, desires, and thoughts. And finally, I conclude that the objections do not show the Transparency Model to be fundamentally flawed, but that it is, nonetheless, limited in
many respects, even if it can fill in some of the gaps left by the Third Person Model. And so we must look to a third model of introspection for help filling in the remaining gaps in a complete account of introspection.

**The Motivation for the Transparency Model**

The Transparency Model is currently defended only by Alex Byrne, but it has also been defended in the past by Dretske (1994), Gallois (1996), and Fernandez (2003).\(^\text{10}\) To keep things simple and fair to the three different presentations of essentially the same view, as well as for substantive reasons that will emerge shortly, I will stick to Byrne’s (2005) presentation of the view, whatever departures from Gallois’, Dretske’s, or Fernandez’ formulations this might entail.

The Transparency Model is motivated by Gareth Evans’ (1982) case of deliberating about the possibility of a third world war. Evan’s observation is this: if someone asks me, “Do you believe there will be a third world war?”, which is a question about my mental states, I do not thereby turn my attention “inward,” directed at my thoughts, to figure out what I do and do not believe. Instead, I weigh the various geo-political factors I am aware of, and based on those I judge whether there will be a third world war. Once I do that, I report my findings based on my examination of the conditions of the world as if I examined my mental states. In other words, I answer the question concerning my beliefs without directing my attention to my beliefs. And importantly, this is a very natural, commonplace, and appropriate technique for answering similar questions about

\(^{10}\) And possibly Peacocke (1998), but he is clearly very eager to distance his own view from Byrne’s. At the very least, Peacocke is motivated by the same considerations as Byrne, Dretske, and Gallois, though he gives those considerations a very different gloss.
our other beliefs. The lesson seems to be this: in order to figure out that you believe that $p$ (or $not-p$), think about whether $p$, and you will have your answer. This shows that our own beliefs are transparent to us in that they can be read off directly from the external world, without requiring any inner focus; hence the name of the model.

In considering Byrne’s account of introspection, it is important to note just how apt Evans’ observation of our introspective practices is. There is simply no denying that at least sometimes we do become aware of our own beliefs and thoughts by focusing on the relevant external factors. For example, becoming aware of your belief that you have a busy day ahead of you will frequently involve thinking back to seeing your calendar this morning, filled with classes and meetings for today. Similarly, recognizing that you only have a week left before your next trip frequently involves checking the date of your trip and today’s date.

I take Evans’ datum to be uncontroversial, even if somewhat puzzling. So the question – in the relevant literature and in this chapter – is not whether introspection ever takes the form as described by Evans, but rather how to give this observation the appropriate structural explanation, and secondarily, just how ubiquitous this form of introspection is.

Byrne thinks that Evans’ case shows introspection to be nothing more than very simple reasoning. In particular, such reasoning involves using a specific rule of inference that takes you from facts about the world to facts about your own beliefs, via your judgment about what the relevant facts indicate. In Evans’ case, the starting facts are about the geo-political factors that might influence the onset of a third world war (as opposed to whether you believe it will happen); but after considering those facts and making a judgment about what they indicate about the likelihood of a third world war, the
result of this reasoning is a response to the question of whether you believe there will be a third world war. Besides using the specific rule that allows this move, there is nothing different about this sort of reasoning from any other. That is to say that, besides the necessary cognitive encoding of the relevant rule, the Transparency Model does not require any additional cognitive mechanisms or abilities – i.e. cognitive “hardware” – that might not already be there for general-purpose reasoning, which is what makes this model economical.

**The Transparency Model**

The central feature of Byrne’s inferential account is the epistemic rule of inference one needs to use to introspect on this model, which is peculiar in a way that we will examine shortly. Byrne calls this peculiar epistemic rule “BEL” and formulates it as follows: “if \( p \), believe that you believe that \( p \).”\(^{11,12}\) According to him, to introspect our beliefs is just to use this epistemic rule.

What makes BEL peculiar and interesting as a rule of inference is that it is subject sensitive in a way most rules of inference, like modus ponens, are not – only the person currently employing BEL is licensed to its conclusion about themselves. In particular, BEL should not be confused with another potential rule (BEL-S): “if \( p \), believe that S believes that \( p \),” for any person S. This rule, though very similar to BEL, is bad in a way

---

\(^{11}\) Andre Gallois, who also defended this view in his (1996), describes introspection in terms of running an argument from \( p \) to the conclusion that I believe that \( p \). It is, however, more common to formulate epistemic rules, as BEL is, as a conditional.

\(^{12}\) There is also a question of how best to formulate an epistemic rule – as in imperative (as BEL is currently formulated) or as a normative proposition. This is a complex debate, but taking a side in it would not influence any of my central arguments for pluralism, so I will set this debate aside, and follow Byrne in formulating BEL as an imperative.
BEL is not, in that BEL-S involves an inference from a state of affairs in the world to a mental state representing that state of affair, without any subject-indexing. For example, the fact that there are six ducks currently on some pond in England gives me no license to conclude that my neighbor believes that there are six ducks on some pond in England. However, BEL, as originally stated, has no such problem since when I concede to the (alleged) fact that there are six ducks on the pond, I make a judgment to that effect. And when I do that, I form, or bring to consciousness, or make available, my belief that there are six ducks on the pond. That is, if you are in a position to use BEL for yourself, then you are automatically considering whether $p$, which leads you to make the judgment that $p$. And once the judgment that $p$ is made, then you have arrived at the belief that $p$, which is also occurrent or conscious. And that does indeed license the conclusion that you have the belief that $p$.

It is also worth noting that in order to use BEL correctly one does not actually need to be a good judge of whether $p$ is true – one simply needs to make that judgment, however epistemically irresponsibly, to secure the result that one has the belief that $p$. So despite the apparent epistemic irresponsibility this rule allows for, it is truth-conducive after all because it is subject sensitive, meaning that whenever someone is in a position to follow BEL by pondering the antecedent, they bring into their consciousness the belief that $p$, which guarantees the truth of the conclusion.

The subject sensitivity of BEL just described is the key aspect in explaining why BEL might be useful at all. In particular, one might worry that BEL does not exemplify a valid inference (in the “necessarily truth-preserving” sense), which would bring into question the sense in which BEL is a good epistemic rule. Byrne describes BEL as self-verifying,
but it is still remains unclear how this label relates to validity, and thereby explains the
goodness of BEL.

To get a better grip on the sense in which self-verification makes BEL a good rule, I
find it helpful to consider Burge’s and Stalnaker’s (independent and distinct)
contributions on the subject. Burge (1996) uses a similar notion to describe our access to
a certain class of judgments, like “I am thinking that there are physical entities” and “I
judge, herewith, that there are physical entities.” He described them as contextually self-
verifying, by which he means that “once one makes the judgment, or indeed just engages
in the thought, one makes it true” (92). Importantly, Burge continued, to labeling
judgments as ‘contextually self-verifying’ is “to remark on their truth conditions, not on
our justification or epistemic warrant in thinking them” (93).

The phenomenon described by Burge illustrates quite well the peculiarity that BEL
has. Although BEL itself is not self-verifying in Burge’s sense, its application – and in
particular, the transition from the subject’s assent to \( p \) to her recognition that she has the
belief that \( p \) – does approximate Burge’s self-verification, in that it relies on the judgment
verifying the belief. What is worth emphasizing, as Burge did, is the role the context
plays in making BEL a good rule, even if not clearly (logically) valid.

This distinction between a logically valid inference and a strictly invalid but
nonetheless extremely compelling inference was also made by Stalnaker (1975). To use
Stalnaker’s example, the following argument is not clearly logically valid but is
nonetheless extremely compelling: “Either the butler or the gardener did it. Therefore, if
the butler didn’t do it, the gardener did” (63). The problem with explaining the
compelling nature of this argument simply in terms of validity is that one would then have to commit to the following argument being valid: “The butler did it; therefore, if he didn’t, the gardener did.” (63) The premise of this much less plausible argument implies the premise of the first, compelling argument; thus if the first argument is valid, so is the second one – which is a counter-intuitive result. Nonetheless, even if the first argument may not be strictly valid, its compelling nature is explanation worthy, in a way that distinguishes the first, compelling argument from the second, implausible one.

Stalnaker’s solution is to introduce a new term, *reasonable inference*, to describe a compelling inference, such as the first one given, when made as a speech act, contrasted with purely semantic relations between the contents of such speech acts, which the familiar notions of entailment and validity concern. The crucial aspect of a reasonable inference is the important role the context of a given speech act plays in determining the relevant information for the given inference. The context of a speech act is thought to place constraints that can change quickly with the conversation, and these constraints determine permissible inferences in a way that a purely semantic notion cannot track. Thus an inference is *reasonable* iff “in every context in which the premises could appropriately be asserted or supposed, it is impossible for anyone to accept the premises without committing himself to the conclusion” (65). Although this definition itself does not specify why the conclusion must be accepted if the premises are, presumably the act of accepting the conclusion is somehow inseparable from the act of accepting the

---

13 Although Stalnaker does not set this definition against the backdrop of being a normal agent, with all the expected reasoning capacities and concepts in place, I assume that such caveats are in order, since the notion of a reasonable inference is a pragmatic one, and not a semantic one.
premises, which can ground the explanation of why reasonable inferences are in some sense epistemically good, even if not strictly valid. To emphasize, this epistemic goodness, whatever it amounts to, cannot be explained strictly in terms of the semantic relations between the propositions involved.

The description of a reasonable inference fits BEL almost perfectly, in that the context – one large part of which is the subject, or the user, of BEL – plays a big role in explaining why BEL is so compelling and apparently truth-conducive, even though it is not strictly valid. One small disanalogy is that Stalnaker described reasonable inferences as being speech acts, as part of a conversation, while the use of BEL does not require audible verbalization, or the larger context of a conversation. However, I see no problem with extending the notion of reasonable inference to BEL, since all the said notion relies on is an explanatory gap between the semantically captured features of an inference and the compelling nature of the same inference, which BEL exhibits. Besides, Stalnaker’s own examples of reasonable inferences are inference types, which appear compelling without being verbalized, and remain so when imagined outside of something that would normally count as a conversation. So I find there is a sufficiently well explained and established notion of inferential goodness, besides validity, that fits BEL, and removes any further need for justifying our use of it.

Nonetheless, granting that BEL is a good rule in some sense, it is helpful to get clear on which way it is not like other, traditionally good (i.e. valid) forms of reasoning, such as modus ponens. Let us consider three differences between BEL, a (merely) reasonable inference, and MP, a (strictly) valid inference – besides the subject sensitivity of BEL discussed earlier. The goal is to illustrate that, even though BEL might not be an
ordinary rule of reasoning, it should not make us more suspicious of it being a rule that we actually could (and do) use.

First, let us focus on the structure of the rules themselves:

\[(\text{MP}) \; p, \text{ if } p \text{ then } q \rightarrow \text{Bel } q^{14}\]

\[(\text{BEL}) \; p \rightarrow \text{Bel } Bel \; p\]

where “Bel” stands for the instruction to believe the relevant contents, and those contents are italicized and in bold.

One disanalogy between these two rules (besides the subject sensitivity of BEL) is that for MP, the sheer truth of the antecedent \((p, \text{ if } p \text{ then } q)\) raises the probability of the (italicized and bold) content of the belief in the consequent \((q)\). However, that is not so with BEL – for example, the sheer fact that there are six ducks on some ponds in England does nothing to raise the probability of my having a belief to that effect, which is the content of the second-order belief I should form by using BEL. This feature of BEL is easily accommodated if we think of it as a reasonable inference, since as Stalnaker described it, part of the definition of a reasonable inference is that its compelling nature is not explained simply by looking at the relations between the propositions involved. So this feature of BEL is as expected.

Another disanalogy between MP and BEL is that unlike BEL, MP can be used within the scope of supposition. That is, you can entertain the possibility of the truth of the antecedent of MP, and thereby form an actual belief that \(q\) under that supposition. But no

---

14 Modus ponens is here formulated to emphasize its role as an epistemic rule, as opposed to its more common role as a deductive rule of inference, which is when it is formulated as a rule schema, and not as a conditional.
such belief under supposition is available when using BEL, because merely supposing $p$
to be true does not lead to the judgment that $p$, and so BEL never gets off the ground.
There might be other inferences that have this feature, though this is a controversial topic.
For example, from the literature on vagueness, the rule that allows the inference from ‘A’
to ‘it is determinately the case that A’ seems to run into trouble when used under
supposition, such as in *reductio* arguments and in conditional proofs, leading to the
conclusion that every claim is either determinately true or determinately false – and
therefore, that there are no vague claims (Heck 1998). Whether this is a problem for the
said rule, or for the generality of reasoning under supposition is debatable; but at least
there are other rules, besides BEL, that enjoy a fair amount of plausibility but not within
the scope of supposition.

And finally, MP preserves the justificatory status of the antecedent and transmits it to
the belief in the consequent. That is, if one is unjustified in believing ‘$p$’ and ‘if $p$ then $q$’,
the belief that $q$ will be likewise unjustified. But when it comes to BEL, whether
believing the antecedent ($p$) is justified has no bearing on the justificatory status of the
second-order belief in the consequent (believing that I believe that $p$). Forming the
judgment that $p$, however epistemically irresponsibly, is sufficient to justify the self-
ascription of the belief that $p$.

In fact, this last feature of BEL is not just one that MP lacks, but is likely a feature
that is entirely unique to BEL as a (candidate) rule of reasoning. As such, it could be seen
as a problem en route to BEL qualifying as a rule of reasoning. In response, there are two
related points to make. First, *every* rule of reasoning is unique in one respect or another,
or else they would all reduce to a single rule. So the sheer fact of BEL possessing a
unique feature should not, by itself, bring into question its status as a (potential) rule of reasoning. And second, this particular feature of BEL is anticipated, given that it is a rule for introspection. Byrne himself thinks that it is this ability of BEL to generate justified beliefs so easily that makes BEL a particularly secure epistemic path to belief self-ascriptions. And since that is what we hope for when constructing introspective accounts, this feature of BEL is predictably unique, insofar as it is the only rule of reasoning meant exclusively for introspection.

With the Transparency Model now on the table, it is helpful to consider ways in which it fits the mold of reasoning-based models, which I outlined in an earlier chapter. In fact, the Transparency Model possesses all four of the features I suggested reasoning-based models have: First, Byrne’s view does not require a dedicated cognitive mechanism – unless we count having to store BEL as one – in addition to whatever reasoning and conceptual capacities a normal agent has. Second, the judgment that \( p \) involved in following BEL clearly is based on (if made epistemically responsibly) and results in contentful mental states (I will say more about this in the next section). Third, it is clear that BEL can be followed consciously and on the person-level. And finally, following BEL is an epistemically significant task, just as all reasoning is, in the sense that failure to follow the relevant rule correctly results in an unjustified belief and the blame is placed squarely on the agent. So I take the Transparency Model to be highly representative of the class of reasoning-based models, possessing all four of the suggested features that differentiate reasoning-based models from perceptual models.

Now I propose to closer examine the claim that BEL is a rule of reasoning, and an inference rule in particular, since so much seems to depend on it. In addition, this will
provide an opportunity to indulge in a minor digression, in the guise of the slowly emerging epistemic consequences of the Transparency Model.

**The Epistemology of the Transparency Model**

Whether Evans’ case should be properly interpreted as involving specifically an inference, as Byrne does, as opposed to reasoning more generally, or something else entirely, is a matter of some contention; Peacocke (1998), for example, disagrees. The problem with interpreting Evans’ case is involving an inference is due to what seems to be a common assumption about inference, as expressed by Peacocke, among others, which is that inferences can occur only between *beliefs*. And it is not clear that the starting point in Evans’ case is a belief, but rather a judgment about the possibility of a third world war – and not everyone agrees that these two states or events are of the same type. One take on judgments is that they are a kind of inner speech act – an inner assertion of belief, sometimes as part of the process of forming that belief – and it is not obvious that such events can be the starting point of an inference.

This problem can be solved without debating the nature of judgments, for on further reflection, the crucial assumption that inferences can occur only between beliefs is likely false. To see this, consider how modus ponens is used under supposition: the starting point in such reasoning is not belief either. And if such classic rules of inference can be extended to starting points outside of belief, there is no reason to assume that other inferential moves cannot be likewise extended. So I see no great threat to an inferential interpretation of Evans’ case due to definitional concerns.
Whether Evans’ case should be thought of as an inference, as opposed to a rule of reasoning more generally, is not the trivial point it appears to be on the surface, because it influences the epistemology that the resulting analysis of introspection inherits. One benefit of inferential reasoning, it seems, is that when one is making an inference, the subject takes the content of the starting belief to support the resulting/inferred belief. In other words, the premise in any given inference acts as a reason to accept the conclusion. Perhaps this is not a necessary condition on inference making – though Peacocke seems to think so, but provides no reason to accept this assumption – but it certainly seems to hold in many cases. And yet, in the Evans’ case, the relevant content, i.e. the proposition (as opposed to the judgment) “there will be a third world war”, does not act as one’s reason for believing that one believes there will be a third world war. And if even we were to call whatever is happening in Evans’ case ‘an inference’, even if an unusual one, the fact remains that it does not seem that the person making such an inference has a reason for her conclusion, in the same sense as premises of many familiar inferences provide one. That is, describing Evans’ case as involving an inference does not change the fact that, from a first-person point of view, no reason is given that might justify the conclusion. This means that if we take the conclusion to be justified, we can only explain this only by resorting to accounts of justification that do not naturally fit with reasons-based explanations, such as the inference being truth-conducive. And this is an unwanted result, since it is unusual for what is supposed to be an intuitive and everyday case to require a particular account of justification to explain it.

Fortunately, there is a way for Evans’ case to accommodate a reasons-based epistemology after all. The idea, which is essentially what Peacocke proposes, is to
explain the epistemic goodness of the proposed inference in terms of it acting as a *shorthand* for a longer inference, which clearly does have room for reasons. Instead of inferring directly from the proposition that there will be a third world war to the belief that one believes this to the case, we *could* become aware of our judgment to the effect that the relevant proposition is true, after making that judgment. And becoming aware of this judgment, that there will be a third world war, can act as one’s reason for the relevant belief.

So on this reasons-based Peacocke-style account, the awareness of the judgment *is* the reason for the resulting second-order belief; but it is left open exactly what this awareness amounts to and how it is achieved. In particular, whether this awareness has the status of a perception-like seeming or a belief is open to interpretation. Which of these two most obvious candidates one chooses will significantly impact the metaphysics of the resulting view of introspection. The details of how exactly one becomes aware of the relevant judgment can play a significant role in complicating Evans’ case to a degree that only a pluralist can accommodate with ease. In particular, it seems doubtful that we become aware of the relevant judgment via reasoning; at least, Evans’ case does nothing to motivate this interpretation. So if one were trying to defend the Transparency Model as the *sole* account of introspection, while trying to accommodate reasons-based epistemology, the result would likely be counter-intuitive at best, and an error theory at worst. But within a pluralist framework, there is no pressure to explain one’s awareness of the judgment through the use of reasoning alone, and so the Peacocke-inspired account can benefit from an entirely different explanation of how we become aware of our judgments that relies on the Perceptual Model, for example. Importantly, relying on a
different model to explain why the Transparency Model is epistemically good does not make the Transparency Model superfluous, since the inference in Evans’ case still needs explaining, which is what the Transparency Model does.

However, it is worth emphasizing that the preceding discussion is not meant to be a defense of any of the described options. Rather, the issue is that interpreting Evans’ case as involving an inference seemed to be incompatible with the talk of reasons, which would be quite unusual and suspicious, and might thereby undermine the motivation for the Transparency Model altogether. So all I have aimed to show is that it is still possible to accommodate reasons talk while interpreting Evans’ case to involve an inference, especially within a pluralist framework.

With the Transparency Model now on the table, let us now examine some objections to it.

**Objections to the Transparency Model**

Brie Gertler (2011) has made several objections to Byrne’s Transparency Model. Her objections involve the distinction between explicit and implicit beliefs, explicit beliefs being the more familiar category, in which propositions are encoded and stored, ready to be retrieved at will, while the implicit beliefs are those that are not so stored but can be easily inferred from the explicitly stored beliefs. It is helpful to use Gertler’s objections to get clear on exactly what the Transparency Model grants us access to.

Gertler’s first objection questions whether the Transparency Model grants us consistent access to explicit beliefs. On her conception of introspection, it is not a process by which one should *generate* (new) beliefs, but only *report* the existing ones. For
example, if at this particular moment I do not believe that I have an overdue library book, and all I do in the next moment is consider whether I believe myself to have an overdue library book, I should *not* self-attribute the belief that I *have* such a book. More generally, Gertler seems to think of introspection as conforming to the following condition (among others): if at a particular time I do *not* have an occurrent, explicit belief that $p$, should I wonder whether I occurrently believe that $p$ at that time, I ought not self-attribute the occurrent belief that $p$ a moment later (if all I do in that moment is introspect).

But the Transparency Model seems to violate this condition, for it is crucial to the proper application of BEL that you (spontaneously) form the judgment that $p$ (or $\neg p$). And this judgment could very well lead (or amount to, depending on what you think judgments are) to a spontaneous creation of a previously non-existent explicit belief, which seems more akin to generating a belief, rather than reporting a pre-existing one. If we consider Evans’ original case in particular, the violation of the Gertler’s condition seems inevitable, since on the natural reading of the case, the reason why you are considering geo-political factors in the first place is because you do not yet have a belief about the likelihood of a third world war. So instead of retrieving a pre-existing explicit belief, you are invited to form a (new) judgment on the relevant topic by the Transparency Model. In short, Gertler grants that this model allows for knowledgeable self-ascriptions of belief, but she is skeptical whether what we so self-ascribe is necessarily a pre-existing explicit belief, and the term “introspection” should be reserved for those processes that carry that guarantee.

In response, it is worth untangling two distinct problems Gertler could be raising with her claim that the Transparency Model violates her condition. On one interpretation,
Gertler could be suggesting that it is a conceptual truth about introspection that it does not, ever, deliver explicit beliefs that did not exist a moment ago. This suggestion, however, is surely false; one of the most central examples of introspective access is our ability to know what we are thinking right now, regardless of whether we have ever entertained those thoughts before. It is not uncommon to form an explicit belief and be aware of this as it is happening. For example, having just looked outside, I have formed the explicit belief that it is very windy out, and am now reporting this in full awareness of said belief. But such cases apparently violate Gertler’s proposed condition, if we interpret it as a conceptual truth about introspection, which speaks against the proposed condition, and not against the Transparency Model.

However, on a different, and perhaps less obvious, interpretation of Gertler’s objection, her suggestion is that, as a matter of fact, we seem to be able to tell (introspectively) whether a given explicit belief is pre-existing, and sometimes even for about how long, or newly formed. For example, my belief that it is very windy out seems to be newer than my belief that Obama is the president, and I seem to recognize this introspectively. This is simply an alleged datum about our introspective powers that an account of introspection should accommodate. But since the Transparency Model invites us to form a (potentially new) judgment every time we use BEL, it does not seem like we can have access to this sort of information, concerning the relative length of existence among our beliefs. In a way, forming the judgment concerning \( p \) each time “screens off” our history with the belief, treating the issue of \( p \) as if it was new each time. And this might be a misrepresentation of our actual introspective powers, and a failure to accommodate a datum about introspection.
A defender of the Transparency Model can make three responses to this interpretation of Gertler’s objection: (a) deny the alleged datum about our introspective powers, and defend the Transparency Model as is; (b) soften the blow by minimizing the set of cases to which Gertler’s suggestion applies, and thereby defend the Transparency Models with some caveats; or (c) bite the bullet by accepting the datum and its ramifications at face value, and thus limit the applicability of the Transparency Model. My own inclination is to accept the last option and thus limit the Transparency Model, a choice that is best defended by comparing the three options in greater detail.

First, there are a couple of ways to deny the alleged datum and resist the suggestion that we can (ever) tell approximately how long we have had a given belief for. One way is to turn the Transparency Model into a kind of error theory, on which such estimations about our history with a given belief are just seemings, a spontaneous belief we form that is probably based on nothing – and the Transparency Model is really good at explaining how we might gain access to those beliefs (after all, the epistemic responsibility with which a belief is or is not formed does not factor into our use of BEL). This, however, is the beginning of a slippery slope towards skepticism about introspection, and as such should be treated as a last resort. A different way to deny the datum is to explain it in terms of access to memory, and then deny that memory retrieval is an introspective process. I imagine this is an acceptable and intuitive response to some, but I prefer not to make such conceptual restrictions before there is a real, independently case-driven need for it.

With regard to option (b) – softening the blow – it is worth noting that the “grain” with which we can compare the relative history of our beliefs is far from fine. In other
words, we can only really compare with confidence (which I will use to track accuracy here, since there isn’t an independent accuracy measure) beliefs that we have just witnessed ourselves form with those that we did not. Knowing that I just formed the belief about it being very windy out, I can pick any other belief of mine, and infer that it is older, helping myself to the fact I did not just detect myself form it. On this story, we would be very bad at comparing the relative age of two beliefs, both of which we have stored for a while; and this does not sound entirely implausible. It seems likely that many of our assessments of the age of our beliefs have to do with largely with the cues from their contents, such as the Obama-being-president belief being formed in 2008, since that is when the election was. If that is how we in fact compare the ages of our beliefs, then the Transparency Model can explain all that, since this is essentially a reasoning-based process. However, I find it implausible that this is the entire story, since not all contents provide such cues, and in any case, older beliefs simply feel different. And this feel is not tracked by the frequency of use, which is another related fact about beliefs that we seem to have access to, and which the Transparency Model has a hard time explaining. Thus I find this approach to Gertler’s objection to be the least plausible.

Finally, one might simply bite the bullet on this matter, as I choose to do, and agree that the Transparency Model does not provide us with all the information about our beliefs, their age (and their frequency of use) in particular. The benefit of this response is that it allows us to continue taking seriously the range of our introspective powers. However, it seems to be a fundamental shortcoming of the Transparency Model as a reasoning-based model that it cannot provide any additional information about our beliefs beyond their content; reasoning-based modes (as I suggested earlier) can only access
contentful states, and the fact that they can only access the content itself, and nothing else, would provide a good explanation of this limit. Of course, this line of response is open only to someone who does not wish to defend the Transparency Model as the sole account of introspection, i.e. a pluralist, which is, of course, my position.

Another objection due to Gertler focuses on the extent to which the Transparency Model can grant us access to *implicit* beliefs while staying true to its motivating idea. Gertler seems to think that what ought to count as a pre-existing implicit belief is a proposition that can be easily deduced only from the pre-existing explicit beliefs, without any new information entering into the relevant reasoning. However, if we are to follow BEL, we would have to investigate or deliberate whether \( p \) each time in order to form a judgment about \( p \). But investigating whether \( p \) each time involves evidence gathering – and importantly, some of that evidence regarding \( p \) will be new, which goes against Gertler’s apparent definition of pre-existing implicit beliefs.

Gertler’s diagnosis of the trouble for the Transparency Model is that BEL instructs us to consider the evidence regarding \( p \) indiscriminately, and so the result of using BEL is not guaranteed to be a *pre-existing* implicit belief. But if we try to fix this aspect of BEL by excluding new evidence from our deliberation concerning \( p \), that would be neither what Byrne’s procedure actually involves, nor is this an advisable modification since we will have to involve introspection *other than* BEL use in order to determine which evidence is old and which is new.

This last step of the objection is an interesting one, and so worth expanding on. The claim that restricting one’s evidence to old only would require additional introspection
without the use of BEL can be defended as follows: in order to guarantee that we rely only on old evidence, we need to be able to access old evidence directly, or at least be able to tell it apart from the new evidence we acquire. And being able to tell whether the evidence is new or old requires a comparison against our existing body of evidence, which is nothing other than introspection. However, the type of introspection required here does not seem like further BEL use; it seems more like direct detecting of our previous evidence, rather than rule- or reasoning-based access. So, the Transparency Model, all by itself, cannot explain our access to the pre-existing implicit beliefs.

Assuming this is the correct way to expand on Gertler’s objection, it illustrates the very same point as her first objection, namely, that the Transparency Model cannot explain how we can apparently differentiate old beliefs from newly formed ones. This is consistent with my earlier claim that reasoning-based models likely can only access the contents of mental states, and beliefs in particular, and not any of their other features. Importantly, this objection does not show that the Transparency Model cannot grant us access to implicit beliefs at all. In fact, the Transparency Model is most naturally understood as a model of access to implicit beliefs – which is what one might interpret Evans’ original case to illustrate. Perhaps this fact can be seen as objectionable in and of itself; after all, the notion of an implicit belief is a highly controversial one. In any case, what this objection does show is that the Transparency Model likely cannot provide the entire explanation of our introspective abilities, which is provides further evidence for the limit of reasoning-based models, and thereby supports pluralism.
Even if Gertler’s objections considered thus far\(^{15}\) do not show the Transparency Model to be fundamentally flawed, it is worth emphasizing one lesson to take away from their discussion, which is that the Transparency Model can only deliver the *contents* of our current beliefs (and perhaps also the fact that they are current and beliefs), but not any other features of beliefs we apparently have access to, such as the history of their existence. And it is very plausible that at least in some cases we do have a general idea of whether we are accessing a previously held belief, or whether we are spontaneously forming a new one. So the objections we have considered bring out an important limit of the Transparency Model, which plausibly exemplifies one of the shortcomings of the reasoning-based models in general, namely, that they can only access contentful states. However, it is also worth noting (again) that perhaps some of the classic cases of introspection do involve accessing our current beliefs – as, for example, while running the *cogito* – for which the temporal history of the belief is not necessary. And that is precisely what the Transparency Model delivers best – the contents of our current beliefs. So whatever the limits of the Transparency Model, its strength should not be underrated.

Besides Gertler’s objections, which bring to our attention the limits of the Transparency Model with respect to the temporal features of our beliefs (i.e. whether they are pre-existing, and for how long, or entirely new), there are other limits of this model worth mentioning. Although beliefs are mental states that most philosophers accept to

---

\(^{15}\) There is a third, thoroughly developed objection to the Transparency Model in Gertler (2011) but it rests fundamentally on a case, which I think is ultimately misdescribed. It involves someone who seems to be plainly irrational – and perhaps self-deceived, on a more sinister interpretation – and yet Gertler analyzes the scenario as if all the usual epistemic notions, particularly those of *belief* and *dispositional belief*, apply to this person without any complications. So I will set this lengthy objection aside.
have no proprietary phenomenology, in the sense that beliefs have sufficient phenomenology to uniquely individuate them, there is still some phenomenology associated with beliefs that does exist and is accessible. This is the phenomenology of the strength of one’s conviction in one’s beliefs. For example, I strongly believe that I will live to 30 years of age, but I am much less sure that I will live to 90 – and I gage my confidence in both of these cases solely on the phenomenology of these beliefs. Importantly, the claim here is independent from the technical epistemic view about the existence of credences. Rather, all I claim is that in many cases we can compare our beliefs, even if very roughly, in terms of the strength of our commitment to them. This claim has to be quite uncontroversial, since it is not only coherent but actually plays a role in the process of updating our beliefs. For example, should we discover a pair of our beliefs to be inconsistent, we give up one of them – and the decision concerning which belief to give up rests largely in our ability to determine which belief we are less confident in.

A natural response to this objection is to read off the strength of one’s conviction from the strength of one’s evidence for p, which does not obviously require anything above and beyond BEL use. However, even if this is how we update our beliefs some of the time, it cannot be the complete story, since it is possible (and perhaps even common) to have unreasonably held beliefs, or at least to have beliefs, which are held more/less confidently than the evidence allows. In particular, all cases of wishful thinking exhibit this feature: I might have no evidence that my latest paper is any good (in fact, I may have evidence to the contrary), but being so tired of rethinking and revising it, I convinced myself that it is in fact in good shape. This shows that we (at least sometimes,
or in part) gage the strength of our beliefs in a purely phenomenological manner, with no regard to evidence supporting it, which is something that the Transparency Model cannot accommodate, since on a plausible understanding of phenomenology, it does not itself have content, which a reasoning-based model can access. It is important to keep in mind, though, that this limit on the Transparency Model does not undermine the proposed procedure as such. Instead, this is just another way in which the Transparency Model is incomplete as an account of introspection.

On the topic of phenomenology, there is one more shortcoming of the Transparency Model worth mentioning. Though I have no knockdown argument for this claim, it is very implausible that we use the Transparency Model all that frequently, or at all, when accessing the beliefs we are currently entertaining. The reason for this claim is broadly phenomenological: in order to employ BEL, we have to first investigate or deliberate on the relevant topic of \( p \). However, when it comes to figuring out what we are currently thinking about, we do not scan through options – “Am I thinking about \( p \)? \( q \)? \( r \)?” etc. – until we reach the topic of our current thoughts. When it comes to occurrent beliefs, we access them without any mental probing of the relevant topics; we can “just tell” what we are thinking about without waiting for a prompt. So from a purely descriptive, broadly phenomenological standpoint, the Transparency Model does not match the introspective process when it comes to occurrent beliefs. It is, of course, entirely possible that this descriptive point might receive the analysis the Transparency model suggests; we would just have to settle for an error theory, in some sense. However, given the other limits on the Transparency Model, and the availability of a different model of introspection that
can accommodate this descriptive point – namely, the Perceptual Model – there is no reason to settle for an error theory in this case.

**The Status and Scope of the Transparency Model**

For all the shortcomings of the Transparency Model, even when beliefs alone are concerned, it is worth pointing to one of its strengths, which is that despite being a reasoning-based model, like the Third Person Model, it does not suffer from a resolution or “grain” problem when it comes to the content of the introspected belief. Recall that the Third Person Model, which uses as input one’s own behavior, can only output rather imprecise contents of mental states. For example, by using the Third Person Model I cannot figure out whether my intention is to be at least 15 minutes early for a meeting or simply not to be late for it. By contrast, the Transparency Model never has a parallel problem – it can always deliver the most precise contents of our beliefs, so much so that I think it is the go-to method for figuring out the exact contents of our beliefs.

Though there are no limits on the belief contents the Transparency Model can access, there is one more potential limit on the Transparency Model we have not considered so far, and it is the most significant one. This limit is on the extent to which the Transparency Model can generalize to mental states other than belief. The key to generalizing the Transparency Model is to notice that it relies on making a judgment with regard to some proposition, which then justifies (or “self-verifies”) the appropriate attitudes towards it. As long as judgment-making is an appropriate way to confirm one’s attitude towards some content, the Transparency Model will be able to account for our access to it. In particular, there is some hope in generalizing the Transparency for
thoughts, desires, hopes, and intentions, to name some of the more discussed propositional attitudes.

Access to Thoughts

Consider first Byrne’s (2011a) account of our access to thoughts. Thoughts are meant to be “mental activities like pondering, ruminating, wondering, musing, and daydreaming” (105). I take it the common element among these activities is the entertaining of propositions. Byrne follows Ryle (1949) in observing that at least some of our thought occurs in “inner speech”; and as a result, we can access our thoughts by hearing an inner voice “speak” them. The experience of hearing it is supposed to be exactly like the ordinary cases of hearing something, with the exception that the ‘something’ is our own thoughts; Byrne describes the experience as “eavesdrop[ing] on herself uttering, in the silent Cartesian theater” some sentence (113). Importantly, this account does not require there actually being an inner voice for us to hear, although it is, of course, compatible with there being one. Rather, it relies on our belief that we are hearing one. So the experience of hearing an inner voice speak is a kind of hallucination.16

The way an inner voice is supposed to explain how we access our thoughts is by allowing us to follow the rule (THINK): “If the inner voice speaks about \( x \), believe that you are thinking about \( x \)” (117). Likewise, if we want to know about the particular

16 Although Byrne himself uses the term “illusion” to describe our experience of the inner voice, it seems more accurate to describe it as being a hallucination. The term “illusion,” as used in philosophy of perception, is usually reserved for misperceiving a property of an existing object, while “hallucination” is used to describe an experience of something that does not exist at all, a total misperception. Considering that Byrne does not postulate that there really is an inner voice, the term “hallucination” seems more appropriate.
contents of our thoughts, rather than its generic subject-matter, we use the rule (THINK-THAT): “If the inner voice says that \( p \) and \( p \), believe that you are thinking that \( p \)” (121). Byrne even takes into account the possibility of imagistic thought, in which case we access them by following the rule (THINK-IMAGE): “If the inner picture is about \( x \), believe that you are thinking about \( x \)” (121).

There are two arguments I will rehearse against Byrne’s THINK-based account of introspective access to thoughts. The first argument is in the form of a dilemma, which concerns the exact interpretation of the central claim that we hallucinate an inner voice when using THINK. The second argument concerns the empirical status of the claim that we access our thoughts with the help of an inner voice. As it happens, there is fairly extensive data on the matter, which can help us determine the plausibility of Byrne’s proposal.

The first objection stems from the claim that our use of THINK is grounded in hallucinating hearing words, which gives rise to a dilemma about the nature of thought. How are we to construe the relation between the hallucination of hearing words and the thoughts that we allegedly introspect through this hallucination? Either these hallucinations are our thoughts, or they are not.

If the hallucinations of hearing thoughts are our thoughts, this is a substantive claim about the nature of thought, which is in itself questionable (for one, why is it a hallucination?), but more importantly it makes THINK uninformative. On this interpretation, THINK is reduced to: “If you are thinking about \( x \), believe that you are
thinking about \( x \)” – while how and whether we are able to follow this rule is left entirely mysterious.

On the other hand, if these hallucinations of hearing words are not our thoughts themselves, the plausibility of THINK is significantly reduced, for without a clear and tight connection between these hallucinations and our thoughts, THINK cannot offer any access to thoughts at all, since there is no thought to have access to.

One way around this dilemma is to construe the THINK rules not as a way of passively hallucinating but as actively imagining words. This interpretation is suggested by Byrne’s response to a potential charge of presupposing an already existing capacity for self-knowledge, which involves the claim that all his account needs is the belief that “inner speech is occurring” whenever one “auditorily imagines words” (117).

The problem with this interpretation of THINK, however, is that imagining does not seem to be a kind of hallucinating. And if the experience grounding the belief about the inner voice is no longer treated as hallucinatory, then there is a genuine question of how we access that experience in order to use THINK – at which point THINK does presuppose the capacity for self-knowledge, and this account becomes explanatorily redundant. Thus it does not seem that inner voice, whether we hallucinate it or imagine it, can ultimately explain how we access our thoughts.

Consider now the second objection. THINK involves a substantive empirical claim, namely, that we access our thoughts with the help of a hallucination of an inner voice. However, as Byrne recognizes, there is some data in (Hurlburt and Akhter 2008), and other work by Hurlburt and his colleagues, that bring into question this empirical claim.
Hurlburt’s research focuses on inner experience, and one of the five types of inner experience he has indentified is ‘unsymbolized thinking’ – a type of thought that is experienced entirely without the medium of symbols, such as words, images, or anything else. Importantly, this type of thought was not deduced from third-person experiments, but rather consistently reported by the subjects themselves, which is significant because it shows unsymbolized thinking to be introspectable.

Briefly, Hurlburt’s claim about unsymbolized thinking comes from years of experimentation using a method he calls “Descriptive Experience Sampling” (“DES” for short, to be distinguished from Byrne’s rule (DES) introduced in the next section). In these experiments, the subjects (who are people without any training in philosophy, cognitive science, or psychology) wear a beeper while they are going about their lives, and once the beeper makes a sound, they are asked to consider their experience at the very moment they heard the beep. They then write down some notes about their experience and report it to a trained experimenter within 24 hours, who helps the subjects get as clear as possible on what that experimental moment was like, exactly, by asking them neutral questions. After years of collecting reports from different subjects about their introspectable experiences, Hurlburt claims to have identified five categories of inner experience: inner speech, inner seeing, unsymbolized thinking, feelings, and sensory awareness. The currently relevant category, unsymbolized thinking, is an experience of a thought, which the subjects persistently deny to be mediated through any symbols, such as images, sounds, words, etc; they are simply aware of the thought itself. Here is an excerpt from an interview with a subject “Evelyn” which illustrates her experience of unsymbolized thinking:
Evelyn: I was sitting on the couch watching TV. On the TV there was a commercial for NetZero. And I was listening to the commercial for NetZero, and thinking about, I wonder how much cheaper that is than Cox Cable? And the pager went off. So it... So as far as I can determine, in my awareness I was holding my coffee mug and y’know kinda wondering to myself, I wonder if Cox... how much cheaper this NetZero could be than Cox Cable. And the pager went off.
Sharon [trained interviewer]: And is that, “I wonder how much cheaper this NetZero is than cable”, is that in your awareness just right at that moment?
E: Um hmm.
S: And is that in words? Or not in words? Or are you saying that? Or thinking that?
E: I was just thinking to myself, I wonder, y’know, if this is actually cheaper.
S: And does that “I wonder if that is actually cheaper”,... So it’s possible to be thinking that in words or not in words, or in pictures, or in... How is that thinking coming to you right at that moment?
E: [Looking powerless: palms turning slightly up, eyebrows raised, voice uncertain] I think just... just thinking about it. Not thinking in pictures or... Just thinking to myself, I wonder if it’s really that much cheaper? Because I keep getting bombarded with commercials for it.
S: And, uh, you’re holding the mug.
E: Um hmm.
S: Is that in your awareness or is that just kind of a fact of the universe: you’re holding it but you’re not paying any attention to holding it?
E: I always pay attention to that mug, because it’s crystal, and I usually use it only on the weekends, and I like the way it feels—it’s real heavy.
S: And so right at this split second, are you noticing the heaviness or the feel or the...?
E: [Returns to the powerless tone of expression] It seems like just the thinking of the Cox Cable versus NetZero is what’s... what I was actually aware of.
(Hurlburt and Akhter 2008:1365, notes omitted)

The absence of an introspectively detectable medium which made the thought accessible to the subject – which is just to say that the thought is unsymbolized – is clear enough in this case, but it is worth making it even more salient by comparing it to other types of experience Hurlburt has identified, in order to eliminate the possibility of assimilating unsymbolized thinking to another type of experience. The following condensed version of the chart from (Heavy and Hurlburt 2008, p. 802) nicely
summarizes and contrasts the important differences between the five types of experience:\(^\text{17}\):

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inner speech</td>
<td>Susan was saying to herself mentally, “I’ve got to get to class.” It seemed just like her voice, but she wasn’t making any sound</td>
</tr>
<tr>
<td>Inner seeing (aka image)</td>
<td>Paul was imaginally seeing the face of his best friend. He could see her neck and her head and that she looked sad, but he could not see anything around her or what she was wearing</td>
</tr>
<tr>
<td>Unsymbolized thinking</td>
<td>Adam was watching two men carry a load of bricks in a construction site. He was wondering whether the men would drop the bricks. This wondering did not involve any symbols, but it was an explicit cognitive process</td>
</tr>
<tr>
<td>Feeling</td>
<td>Courtney was unequivocally angry, although it was difficult for her to describe how this anger presented itself to her. It seemed to be conveyed by or accompanied by a tight feeling in her chest and a little shakiness in her hands, but she could not be definite about those aspects</td>
</tr>
<tr>
<td>Sensory awareness</td>
<td>It was windy, and Harriet was feeling the cold breeze on her left cheek and her hair moving, tickling her forehead. She wasn’t thinking about those aspects, but she was explicitly noticing the coldness and the moving hair</td>
</tr>
</tbody>
</table>

There is little doubt that, at least introspectively, the experience of unsymbolized thinking is distinct from the other types of experience. The problem that the case of

\(^\text{17}\) I am omitting a column from the chart in the original source that describes each type of experience, since for our purposes, only the examples of the different types of experience matter.
unsymbolized thinking poses for Byrne’s account of introspective access to thoughts is that it prevents THINK together with its variants, THINK-THAT and THINK-IMAGE, from explaining our access to all thoughts. If some thoughts are unsymbolized, there is apparently nothing that can act as an antecedent to a potential THINK-like rule. So if any thinking is unsymbolized, our access to it cannot be explained not just by THINK specifically, but also by any other rule that would fit the pattern of the Transparency Model.

One natural response to this problem, which Byrne in fact formulates, is to doubt whether any thinking is truly unsymbolized, particularly in light of Hurlburt’s admission that the “apprehension of an unsymbolized thought may involve the apprehension of some sensory bits, so long as those sensory bits are not organized into a coherent, central, thematized sensory awareness” (2011, 149-150). Perhaps such “sensory bits” could be enough to reconstruct the speech of the inner voice, and allow THINK to proceed as specified.

However, this response is inconsistent with Hurlburt’s data, since none of the subjects experienced anything like “sensory bits” at an introspectable level, from which they could reconstruct their thoughts. Whatever the “sensory bits” are, they are meant to be insufficient for a full reconstruction of thought. So it appears that even if THINK were the correct account of how we access some thoughts – those that present themselves as spoken by an inner voice – there is no rule that can explain our access to all thought.
Thus I doubt that THINK can explain our access to thoughts, and even if it can, there is likely no transparent rule of reasoning that can explain our access to all thoughts. So the Transparency Model fails to generalize completely to all thought.

**Access to Desires**

Perhaps generalizing the Transparency Model to desires is more fruitful. Byrne (2011b) suggests that accessing one’s own desires is a matter of following the rule DES: “If φing is a desirable option, believe that you want to φ” (177).18 There are two crucial clarifications Byrne makes with regard to this rule. One is that this rule is meant to be non-circular in the sense that for an option to be judged desirable is not necessarily for it be judged desired. So desirability and actually being desired are meant to come apart in a substantive way. But – and this is the other crucial point – being judged desirable is not identified with being judged to be the best option either, “best option” being the most reasonable option, all things considered. Rather, Byrne opts for the *Oxford English Dictionary* analysis of desirability, namely, possessing “the qualities which cause a thing to be desired: Pleasant, delectable, choice, excellent, goodly” (176).

Of the transparent rules of reasoning considered so far, DES is the first one to be formulated in a way that is potentially circular, in the sense that the rule itself relies on our antecedent grasp on and independent access to desires. BEL, for example, is clearly non-circular; it instructs us to judge the truth of p, without inviting us to consider our beliefs directly. By contrast, DES does appear to require us to track our desires directly, and not via some independent (though of course related) property, like truth in the case of

---

18 To emphasize, this rule DES is to be distinguished from Hurlburt’s experimental method “DES” discussed earlier.
BEL. Either DES is in fact circular, or it is not; either way, it does not provide an acceptable account of our access to desires, or so I will now argue.

First, consider the case in which DES is in fact circular, which would happen if being desirable should turn out to be the same as being desired. There are a couple reasons why this outcome would make DES a poor explanation of our access to desires. On the one hand, if we continue thinking of DES as a rule of reasoning, its circularity would render it uninformative. It is doubtful that we actually employ rules of the format, “If I find myself in the state S, believe that I am in S,” which is what DES is reduced to, on a circular interpretation. Finding yourself to be in some state simply is, or immediately leads to, believing that you are in that state; no reasoning is necessary. On the other hand, we could give up the idea that DES is a rule of reasoning, and reinterpret it instead of a rule that describes the function of some (likely sub-personal) causal, belief-forming mechanism that executes the transition from finding yourself to be in some state to believing that you are in that state. However, this option involves giving up on DES as a version of the Transparency Model, since DES would no longer be thought of either as a rule of reasoning, or as being transparent, given that the antecedent would involve awareness of a mental state, and not some external factors. In short, we would be reinterpreting DES as involving a non-reasoning and apparently dedicated cognitive mechanism, which would turn DES into a version of the Perceptual Model. Thus, if DES is circular, it does not offer an informative, reasoning-based account of introspection of our desires.

Let us now suppose that DES is not, in fact, circular, as Byrne intends it to be read. This is the case if being desirable and being desired come apart in a substantive way.
The problem with this reading of DES is that, unlike BEL, DES would not be self-verifying. That is, unlike with judging that \( p \) (is true) and believing that \( p \), there is no causal link between judging that \( \varphi \)-ing is desirable and desiring that \( \varphi \). Byrne recognizes this, but nonetheless opts to describe DES as being “practically self-verifying” (178). But on this interpretation of DES, it does not offer us a way to track our desires; at best, it can act as a heuristic device that signals the possible presence of the relevant desire. In turn, this is a problem because we do not have a good grasp on the exact (non-causal) relationship between being desirable and being desired, if we take the non-circularity of DES seriously. And without a good grasp on this relationship, we cannot assess the usefulness of DES as a heuristic device. It would become an entirely empirical matter whether, and to what extent, DES can predict the presence of the relevant desires. Moreover, we have no reason to think that DES would be a particularly good predictor of our desires – if, again, we treat being desirable and being desired as genuinely separate and not causally related. All things considered, this route involves a significant reduction of the epistemic responsibility involved in using DES, while reasoning-based models tend to have a high level of epistemic responsibility. And so, once again, the extent to which the resulting view would be a reasoning-based one is questionable.

Thus it seems that neither a circular, nor a non-circular interpretation of DES makes it a satisfying account of our access to desires for reasons that stem from it being a reasoning-based model. In fact, if we take “access” to imply a causal link to whatever is being accessed, DES is not an account of access at all, on either the circular or the non-circular interpretation, unless we treat DES as a perceptual, rather than a reasoning-based, account. Moreover, it does not seem likely that there is room for improving the chances
of DES providing a good account of our access to desires by altering its formulation, since the discussed circularity dilemma concerns the features of DES that are not tied to Byrne’s actual formulation of it. However DES is framed, as long as it treated as a rule of reasoning, it will be either circularly stated or not, and all the problems I have been outlining for it so far will resurface. So I find it unlikely that our access to desires can be explained by a transparent rule of reasoning.

Access to Intentions

Let us now consider Byrne’s recent (2011) analysis of our access to intentions. Modeling his account on the account of access to beliefs described in the previous section, he proposes that one infers that one intends to \( \phi \) from the judgment that one will \( \phi \):

I will \( \phi \)  
I intend to \( \phi \)

Byrne calls this the bouletic schema. This schema exploits the much-discussed proposal that intentions entail (relevant) beliefs, but it does not require the stronger thesis that intentions are a kind of belief (e.g. Velleman 1989).

Furthermore, this schema is supposed to parallel BEL in the following respect: When using BEL, one considers whether \( p \) and subsequently judges that \( p \) (or not-\( p \)), which brings to mind one’s belief that \( p \) (or that not-\( p \)). Similarly, when using the bouletic schema, one judges that one will \( \phi \), thereby brings to mind one’s belief that she will \( \phi \); and when this belief is held without any (or any good) evidence with regard to the likelihood of one’s \( \phi \)-ing, this belief can only signal the presence of one’s intention to \( \phi \), or so Byrne’s account goes.
The caveat about the absence of (good) evidence about the likelihood of \( \phi \)-ing is an important one in that it is meant to block a potentially fatal objection to Byrne’s account. This potentially serious problem for the bouletic schema, which is one that Byrne recognizes, is presented in the form of cases when one does have ample evidence that one will \( \phi \). In particular, there are two types of cases that pose the (prima facie) problem of defeasibility. These are the cases of “foreseen but unintended consequences” (217), and what might be similarly called foreseen but unintended actions. The former case comes from Bennett (1981), in which a bomber only intends to destroy a factory, but also fully expects to be thereby killing ten thousand civilians – an unintended side-effect. The latter case, due to Anscombe (1957), involves a student who expects to fail an exam in virtue of being insufficiently prepared for it, even though this is not an intended result. In both cases, should the agents reason according with the bouletic schema, they would determine themselves to have the relevant intentions, which clearly does not seem to be the case.

Byrne’s solution revolves around the contrast between cases in which we have good evidence that we will \( \phi \) and those in which we do not have such evidence. The bouletic schema is only meant to be used when we have little or no evidence with regard to \( \phi \)-ing. However, when we do have good evidence that we will \( \phi \) – as in the problematic cases above – we simply “will not reason in accord with the bouletic schema” (218). Importantly, this claim does not involve an invitation for a further rule, distinct from the bouletic schema, to accommodate cases in which we attempt to access our intentions when we do have good evidence that we will \( \phi \). Rather, I take Byrne to be suggesting that the question of our intentions simply does not arise in the presence of good evidence for
the likelihood of us φ-ing. This is because when we do have such evidence, we would use
BEL instead, since “from the first-person point of view, an enquiry into one’s evidence is
(near enough) extensionally equivalent to an inquiry into one’s beliefs: one takes P to be
part of one’s evidence just in case one believes that one believes P” (218-9, footnote
omitted). So the bouletic schema is meant to provide an account of access to all
intentions after all, even with the apparent restriction on evidence.

As elegant as this solution is, I believe the bouletic schema is nonetheless open to
further objections, which leave it either unnecessary or insufficient as an account of
access to intentions, depending on how belief-like intentions turn out to be on Byrne’s
view. The remainder of this section is dedicated to defending this disjunctive claim.

First, consider whether the bouletic schema is necessary to explain our access to
intentions. Insofar as there is meant to be a parallel between Byrne’s accounts of access
to beliefs and access to intentions, it is worth thinking about what exactly the relation
between beliefs and intentions is supposed to be on his view. One of the crucial points of
Byrne’s view that offers some insight into the relation between beliefs and intentions on
his view is the last substantive point in the previous section – that when we have ample
good evidence about what will happen, we “will not reason in accord with the bouletic
schema” (2011, 218). Rather, Byrne suggests, using BEL in this situation would be more
fruitful or appropriate. This is a vital clue to reconstructing what Byrne takes to be the
relation between intentions and beliefs, so let us consider it further.

Specifically, it is unclear exactly how to interpret what initially sounds like an
empirical prediction that we “will not reason in accord with the bouletic schema” in light
of good evidence for φ-ing (my emphasis). Is it that such reasoning is unnecessary, or
that it is impossible? If using the bouletic schema when good evidence for φ-ing is available is merely unnecessary but nonetheless possible, then we are still capable of wondering whether we intend to φ, even when we believe we will φ. On this interpretation, the bouletic schema immediately turns out to be insufficient to fully capture our introspective abilities, since it leaves room for us to access our intentions without using the bouletic schema. There is a response available on Byrne’s behalf to this line of reasoning, which I discuss in the next section, dedicated to the insufficiency of the bouletic schema. For now, let us keep our focus on the necessity of the bouletic schema and consider a different interpretation of Byrne’s empirical prediction.

The other way to read the prediction that we “will not” use the bouletic schema when we have good evidence for φ-ing is that reasoning in accord with the bouletic schema is simply impossible in those circumstances. However, if using the bouletic schema when good evidence for φ-ing is available is outright impossible, there needs to be a further explanation of this curious alleged fact. The most obvious explanation of why it would be impossible to reason according to the bouletic schema when we have good evidence for φ-ing is that the connection between beliefs and intentions is closer than Byrne permitted himself to suppose, i.e. closer than mere entailment, and is in fact more akin to Velleman’s treatment of intentions as beliefs. But if intentions are treated as identical with beliefs, there would be no need for a separate account of how we access intentions, since Byrne already has an account of access to beliefs (by using BEL). Thus, if intentions are treated as beliefs, the bouletic schema is unnecessary in an account of access to intentions. This establishes one of the disjuncts of my conclusion.
Given that Byrne does see a need for the bouletic schema in an account of access to intentions, let us now suppose that intentions are not entirely belief-like on his view. However, this supposition leads to the other disjunct of my conclusion about the bouletic schema – that it is insufficient to explain our access to all intentions. In other words, the bouletic schema cannot explain our access to all intentions.

First, let us revisit my earlier suggestion that we can apparently still wonder about our intentions with regard to φ-ing even if we have ample evidence that we will φ. If this is the case, then we can access our intentions without the help of the bouletic schema, which renders it insufficient as an account of introspection of intentions.

To see that we can access our intentions without using the bouletic schema, consider a case from Bratman (1987). In it, one intends to mail the bills on the way to work, but knowing oneself to be forgetful, one does not believe that one will in fact mail the bills. What this case illustrates is that the forgetful subject can apparently access the intention to mail the bills despite the absence of the belief to do so. This suggests that the bouletic schema was not used to access one’s intentions in this case. There are ways to undermine the Bratman case, and to explain away the apparent lack of belief or the apparent presence of the intention, such that the belief and the intention do not really come apart. However, the closer intention and belief are treated, even extensionally, the less obvious the need for the bouletic schema at all, as was argued in the previous section.

A response on Byrne’s behalf to this threat of sufficiency of the bouletic schema could center on the distinction between having good evidence and using the said evidence. In particular, if one has good evidence that one will φ and wonders about whether one really intends to φ, what one could be doing is setting aside the evidence one
has with regard to $\varphi$-ing in the process. In the Bratman case, that would mean for the forgetful person to set aside all the good evidence one has about being forgetful and ask oneself whether one will mail the bills. One might even pretend that there is absolutely nothing else one has to do that day, and that there are notes written all over the house reminding gone to mail the bills. Without taking one’s forgetfulness into account, it seems that this person would sincerely think that the bills will be mailed – which is just the result Byrne would predict upon correct application of the bouletic schema, thus predicting that this person in fact intends to mail the bills, despite being forgetful.

The problem with this response, though, is that once the evidence for $\varphi$-ing is set aside, it does not seem that the belief about $\varphi$-ing can remain either, at least in cases like this one. The entirety of one’s belief about the likelihood of the bills being mailed in the Bratman case comes from being aware of one’s forgetfulness. Without any beliefs about being forgetful, it is unclear that this person would have any beliefs about the likelihood of the bills being mailed. And without a belief about the possibility of $\varphi$-ing – which, in this case, is the likelihood of the bills being mailed – the bouletic schema cannot be used.

More importantly, the point of the Bratman case is to illustrate a general introspective ability, namely, that we can introspectively access our intentions despite our belief that the intended action is likely to remain unactualized. The presence of good evidence that the bills will not be mailed is not meant to be an interfering factor that we may choose to set aside; rather, it is meant to illustrate the power of our introspective ability to detect our intentions in full awareness of their fruitlessness.

Likewise, it is commonplace to make promises like these: “I really don’t know if I will get around to doing this, but I will certainly try!” In other words, sometimes we have
intentions to φ without having any belief about the likelihood of φ-ing at all. And as mentioned earlier, without a belief about φ-ing, the bouletic schema cannot be used.

Thus, if beliefs and intentions come apart, the bouletic schema cannot capture the full extent of our introspective access to intentions in certain circumstances, like having beliefs about their fruitlessness, or having no beliefs at all about their fruitlessness. The bouletic schema is, therefore, insufficient as an account of introspective access to (all) intentions. Alternatively, if intentions are beliefs on Byrne’s view, the bouletic schema is straightforwardly unnecessary even on Byrne’s own view, since he has already provided an account of introspective access to beliefs.

*Hope for Further Generalizations*

The issue of self-attributing motivational states, like desire and intention, is amplified when we consider mental states that have an even greater phenomenological component than desire, like fear and sensations. Even if such mental states turn out to be representational, which is minimally necessary for the Transparency Model to generalize completely since (as a reasoning-based model) it can only access contentful states, it would be wildly inaccurate to describe our access to them as reasoning. In particular, fears and sensations like pain need no self-questioning and subsequent rule application to establish that a particular phenomenology occupies our attention.

---

19 Here I am assuming that sensations are (largely) mental, which is one necessary condition for the Transparency Model to be able to account for our access to such states. Should such states turn out to be (largely) non-mental – as, for example, Hill (2009) argues – the failure of the Transparency Model to generalize to them follows immediately.
Take fear as an example – there is an elegant rule available that captures the vast majority of the situations in which one experiences fear: “If you perceive a threat in \( x \), believe that you fear \( x \)”. When thinking about applying this rule, it is important to distinguish fear from anxiety, which has a similar displeasing phenomenology to fear but can occur without the perception of threat. What this means is that one cannot object to the proposed fear rule by posing the following counter-example: Suppose I am generally afraid of spiders, but I do not experience fear when I see spiders on TV. This case would be a problem if the fear rule relied on actually threatening situations in the antecedent. But since the threat has to be only perceived, such counter-examples do not apply. In fact, I am quite confident that the proposed fear rule is as good as BEL is in the relevant way – in that it seems be equally self-verifying – and let us suppose this is indeed the case. Even so, it is a mistake to think of the Transparency Model as accurately describing introspection in phenomenologically-laden mental states like fear. The reason for this is that the proposed fear rule, and the Transparency Model more generally, paints an unbelievable picture of our access to our fears. Mental states like fears and sensations consist largely (if not wholly) in their phenomenology, which simply “hijacks” our attention without any apparent effort on our part. In fact, it usually takes effort not to notice our fears and pains. And to imagine that this attention-hijacking happens through prompting oneself to see if one is fearful or in pain first, and only then coming to the judgment that one is, and so finding out about one’s current fearful experience or pain is to push the boundaries of the coherent. And yet, fear and pain are among the paradigm examples of mental states that are accessible through introspection. In short,
phenomenology does not seem to be properly (i.e. contentfully) “formatted” for access via reasoning, and the Transparency Model in particular.

More generally, as soon as we start considering cases of compulsion, phobia, and other irrational tendencies, the Transparency Model breaks down because it is only equipped to explain our access to mental states that can exist to us as something beyond paralyzing phenomenology. Whether a rule of reasoning, such as the fear rule proposed above, can capture how that phenomenology arises is only a post hoc matter; our awareness of being in such a state does not come via our self-questioning of whether we are in it. These states simply invade our attention and wait for no deliberation on our part to be discovered. In fact, we do not even need to go as far as pathological examples to see this limit on the Transparency Model – any rational, but strong and phenomenologically-laden mental state will do, even if it has a cognitive component, like hope, or most clearly, a strong emotion or sensation. The phenomenon of one’s attention being hijacked by the immediately recognizable headache is quite familiar, which does not require us wondering about it prior to the fact. And this attention hijacking can only be plausibly explained in terms of a dedicated cognitive mechanism, which fits more naturally with the Perceptual Model, rather than with the Transparency Model.

**Conclusion**

It should be now clear why I am pessimistic about the prospects of generalizing the Transparency Model to mental states other than beliefs – and not even to all aspects of belief. However, my resistance to generalize the Transparency Model might undermine my desire to defend even a limited version of it, tailored to beliefs alone. As Byrne
(2011) argues, if an economical model, like the Transparency Model, accounts only for parts of our self-knowledge, we would expect to observe dissociations in mental state self-attributions across all the mental states. That is, some people could (and would, in a sufficiently large sample) lack the ability to access some mental states but not others. And this empirical prediction seems to be simply false, Byrne thinks, since we do not observe such dissociations in our ability to self-attributive mental states. (This argument is akin to Shoemaker’s “self-blindness” argument, though significantly more modest.)

There are three points to make in response. First, Byrne’s own accounts of our access to thoughts, desires, and intentions are prone to this dissociation worry. Each mental state type seems to require a different rule, even within the scope of the Transparency Model, so it is not clear that there is sufficient continuity within the model (and across mental state types) to guarantee the absence of dissociations. After all, there is no conceptual or even physically necessary reason why someone might have BEL but lack, say, THINK. So if the dissociation objection is a problem at all, it is a problem for everyone, not only to a defender of pluralism or of the Perceptual Model.

Second, the inference from pluralism to the presence of dissociations is too quick. For even if there are multiple mechanisms responsible for our mental state self-attribution, they could still have an important element in common, such as central processing, which would preclude the possibility of dissociations. Moreover, the dissociations in my particular type of pluralism are likely very limited, if not absent, since I propose that the different mechanisms responsible for mental state self-attribution overlap in their application; so if one cognitive mechanism is malfunctioning, a different
cognitive mechanism might nonetheless grant us (partial) access to the same mental state type as the malfunctioning mechanism.

Finally, it is by no means empirically established, or even sufficiently investigated, whether that there are any actual dissociations in our ability to self-attribute different mental states. The claim that there are no such dissociations should seem equally likely to be true as it should seem false at this stage. Few could have thought likely that humans might suffer from a condition in which negative affect and the experience of pain are dissociated, and yet such a condition was discovered (i.e. pain asymbolia). In fact, the beginnings of a parallel dissociation in mental state self-attribution can be seen in Wilson and Nisbett’s seminal (1977) study, which showed the poor accuracy of our access to our cognitive processes – and if we combine that finding with the near infallibility of the Transparency Model in the form of BEL use, we should already expect a dissociation based on the different accuracy levels of our access to belief contents and our cognitive processes. So, it is, perhaps, the burden of the pluralist to establish such dissociations, but it is by no means obvious that they are absent.

In conclusion, and despite the problems with generalizing, the Transparency Model is otherwise a successful account of introspection of beliefs. It is limited in its own ways, as is the Third Person Model, but these limits are only that, and not fatal flaws. However, the limits that the Transparency Mode does possess stem from it being a reasoning-based model of introspection, and so supplementing it with a different reasoning-based model is futile. Thus a perceptual model is needed to fill in the gaps left by reasoning-based models.
Chapter 4: The Perceptual Model of Introspection

In the previous chapters, I have argued that both the Third Person Model and the Transparency Model are necessary in order to account for our access to the different mental states. This already establishes a minimal version of pluralism insofar as two different models have been shown to be necessary. However, this type of pluralism – consisting of two economical models – is neither as exciting as pluralist accounts can be, nor is it sufficient. It is not as exciting as it could be because it consists of economical models only, which make a relatively uncontroversial combination. But more importantly, the two economical models are not sufficient to explain our access to all mental states because, as I have argued, neither the Third Person Model nor the Transparency Model can grant us consistent access to phenomenologically-laden mental states like sensations, emotions, fears, and some desires.

To fill the gap in the current pluralist account left by the Third Person Model and the Transparency Model we need to consider the only model of introspection still left to be discussed – the Perceptual Model. As it will emerge in this chapter, although the Perceptual Model is usually thought to be “extravagant”\(^{20}\), in that it usually involves a cognitive mechanism dedicated to introspection, it is not necessarily so extravagant that it is as controversial as some versions of it have been historically.

In this chapter, I begin by outlining the remaining mental states that we can introspect, and the access to which has not yet been accounted for. Next, I move on to examining the Perceptual Model, including its variants, and the different motivations for

\(^{20}\) To borrow the term from Byrne (2005).
its different versions. Given the wide variety of Perceptual Model variants available in the literature, I focus on three versions of it, which are necessary for a complete account of introspection – a constitutive version, a monitoring version, and an attention-based version. I then defend my preferred versions of the Perceptual Model against some of the more notable objections accumulated over its long existence, and conclude by considering how the Perceptual Model fits into a pluralist account.

**The Task for the Perceptual Model**

First, let us consider what it is that we need the Perceptual Model to accomplish to complete the pluralist account, namely, explain how we are able to access mental states heavy with phenomenology. This is no ordinary task, for two reasons. One is that the description I have been giving to the problematic group of mental states – “heavy with phenomenology” or “phenomenologically-laden” – is not a natural kind, but rather applies to, at least *prima facie*, a very heterogeneous group of mental phenomena. So, again *prima facie*, we cannot expect all the mental states in this heterogeneous group to be accessed in the same fashion.

And second, the issue of access to phenomenology suffers particularly greatly from a tension that comes from considering the problem of ‘knowing our own minds,’ as it is sometimes called. In the debate concerning the amount and security of access to our own minds, it has long been deemed that if we have anything close to infallible access to any aspect of our mental life, it is likely to be the phenomenology, since it is usually defined as ‘however things *seem* to us’. And by definition, we cannot be wrong about how things *seem* to us, so the intuition goes.
Having said that, as with any controversial topic, there are countervailing intuitions and arguments, such as those presented in Hill (2009), which suggest that even the most prototypical phenomenological states like feeling pain may not be infallibly accessed. However, I do not argue that the view of phenomenology just outlined is the correct one; rather that it has enjoyed a good amount of plausibility. Thus, one of my goals in this chapter is simply to make room in a pluralist account for a view on which phenomenology can be infallibly accessed. If one find themselves drawn to a different conception of phenomenology, it is (or rather shortly will be) much clearer how pluralism can accommodate it.

Returning now to the view on which phenomenology can be infallibly accessed, the reason why infallible access can be a problem is that it can only be secured through a constitutive account akin to Naïve Realism about perception: just as according to Naïve Realism the objects of perception partially constitute our experience of them, so must our beliefs about our phenomenology be partially constituted by that phenomenology to guarantee their complete accuracy. But it is commonly thought that such an account would hardly be useful in the more general attempt to make our own minds more accessible to us, for whichever beliefs we could form about our phenomenology would be partially constituted by it, which is fleeting, and would therefore resist integration with the rest of our beliefs. For one, these beliefs partially constituted by phenomenology can only be expected to be as lasting as the phenomenology itself, and so how further beliefs are to be based on them is unclear. Also, though there are representational accounts of

---

21 This is, of course, a different constitutive account from Shoemaker’s: his concerns the constitution of second-order beliefs by first-order beliefs, while this concerns the constitution of beliefs by phenomenology.
phenomenology itself, the more intuitive view is that phenomenology is thoroughly unconceptualized, and so how beliefs could have such contents is also unclear.

Putting these two problems together makes it very hard to see how any account of phenomenological access that incorporates these naïve ideas can help us resist the looming skepticism about knowledge of our own minds.

Though resisting skepticism is not the purpose of this chapter, this problem of phenomenological access is nonetheless highly relevant, for it highlights the greatness of the burden placed on the Perceptual Model. In fact, there is a further dimension to the problems the Perceptual Model faces. Even if the naïve phenomenological picture just discussed is philosophically salvageable, we have to be very careful not to extend the infallibility claim that concerns our phenomenology to other aspects of our mental life. For it would be equally unfortunate (and inaccurate) to be committed to the claim that we have complete and infallible access to everything in our minds. In short, we need to take the middle road when balancing the amount of error in introspection across different mental state types: for some aspects of our mental life (like phenomenology) infallible access might be desirable, but for the rest, which makes up most of our mental activity, there must be room for error in our access to it.

Despite the difficulties outlined for the Perceptual Model, it is the most versatile of all the models we have considered so far, so there is still hope for integrating it into a pluralist account of introspection. Let us now examine this model in detail to see how it can overcome these and other difficulties.

*Varieties of Perceptual Models*
The Perceptual Model offers such an explanatorily powerful account of introspection that it is potentially the only model one would need to explain our access to all of our mental life. Just like our perceptual mechanisms give us all we need to know about our environment, so introspection, when modeled after perception, can give us all we need to know about our own minds. However, it will become apparent in this chapter that relying solely on the Perceptual Model to explain our introspective access to all mental states is not plausible.

The Perceptual Model of introspection comes in many and significantly different varieties, but they all construe introspection as involving a dedicated faculty or mechanism, in some way analogous to perception. And just as our perceptual faculties can accommodate any type of environmental input, so introspection, modeled after perception, can access any type of mental state, though many philosophers like to nonetheless restrict the Perceptual Model to a handful of mental state types. But even with these restrictions, the Perceptual Model is frequently thought to describe introspection of at least those mental states that have strong phenomenology, like sensations, emotions, current perceptual experiences, etc. This is because there is simply no other good candidate explanation of how we can access our phenomenology except via some dedicated cognitive mechanism. However, how reliable that access is, whether this mechanism can also access other aspects of our mental life, and how strong

---

22 There is one exception to this generalization – the constitutive account of phenomenological access, mentioned earlier, which does not require a dedicated cognitive mechanism, and which is why it is as frequently considered to be a perceptual account as not. I will say more about this later in this chapter.

the analogy between introspection and perception ought to be – these are the points of 
departure for the different versions of the Perceptual Model.

The version that is most likely to have been Descartes’ leaves the parallel between 
introspection and perception at its minimum, limited to the idea that both require 
dedicated faculties that can provide us with a body of knowledge. Beyond that, 
introspection is different in every respect from perception on this version of the 
Perceptual Model. Whereas perception is (sometimes thought to be) threatened by a 
“veil” of sense-data, introspection on this view is completely unmediated, granting us 
direct access to our mental states. Because of this direct access, introspection is thought 
to be infallible on this view (a claim that is almost universally rejected now in this 
unrestricted form). Chisholm’s view, which was similar, highlights a different feature of 
introspection (besides its directness), that too can lead to a limited infallibility claim – a 
property of mental states called “self-presentation”\textsuperscript{24}: whenever you are in a given mental 
state and consider whether you are in it, you know, or are at least certain, that you are 
(the basis of this property is left unspecified). Incorporating some version of an 
infallibility claim is the central feature of this version of the Perceptual Model.

However, the more recent defenses of a (distantly) Cartesian view are significant 
departures from the original, and are never left unrestricted, usually limiting the 
applicability of the crucial infallibility claim to phenomenology (e.g. Gertler (2001), 
Papineau (2002), Chalmers (2003), and Horgan and Kriegel (2007)). Let us call such a 
version of the Perceptual Model, the \textit{Constitutive Perceptual Model}. On this view, the

\textsuperscript{24} This seems to be a similar claim to Shoemaker’s (1994) “self-intimation” claim.
phenomenology of a relevant experience is embedded in, or partly constitutes, the introspective belief about that experience. This partial constitution of the belief by the relevant phenomenology allows one to have limited but nonetheless infallible access to that phenomenology.25

I believe such infallible access to phenomenology is best secured by a very simple, dedicated cognitive mechanism that embeds the relevant phenomenology in the introspective belief. Such a mechanism would be too simple to be seriously objectionable on the grounds of requiring too much cognitive space. And since the resulting introspective beliefs would be demonstrative and partly constituted by the relevant phenomenology, there is no threat of pervasive error that usually accompanies contingently formed beliefs, thus securing the limited infallibility claim. Moreover, the need for such a dedicated mechanism would make it obviously a version of the Perceptual Model, since reasoning-based models tend to avoid dedicated cognitive mechanisms.

However, it may be possible to construct this view without a dedicated cognitive mechanism, and instead relying on whatever mechanism is responsible for forming demonstrative beliefs more generally, since the introspective beliefs one forms on this view are demonstrative. In fact, it is unclear to me whether the aforementioned recent defenses of this type of view advocate the use of any dedicated mechanism of introspection in order to access our phenomenology. But one might worry that without a

25 It is worth highlighting that this view is different from Shoemaker’s Constitutivism, according to which second-order beliefs are partly constituted by first-order beliefs. On this view, the constitution claim does not involve two kinds of beliefs, but rather beliefs and phenomenology.
dedicated cognitive mechanism, this view is not clearly perceptual. Although the issue of proper labeling is hardly of great importance, it is worth noting a parallel between this view of introspection and Naïve Realism about perception. Both views involve a constitutive relation between the resulting beliefs and the objects of those beliefs. Moreover, both views can be seen as presenting a route to avoid skepticism of the relevant sort (skepticism about the internal world in the case of Constitutive Perceptual Model, and skepticism about the external world in the case of Naïve Realism). I consider this parallel between a clearly perceptual account and the Constitutive Perceptual Model to be sufficient reason to consider the Constitutive Perceptual Model to be appropriately labeled as “perceptual”.

Moving on to the other extreme of Perceptual Models, introspection is sometimes thought to have almost everything in common with perception. David Armstrong (1968), for example, claims that introspection is literally a “self-scanning process” (419) that “scans” (or monitors) one’s mental states, and which results in higher-order awareness of those states. Besides taking in information about different types of things, external objects in the case of perception and mental states in the case of introspection, there is no metaphysically or epistemically significant difference between perception and introspection for Armstrong. On his view, the only significant difference between perception and introspection is just the distinctness of the mechanisms that underlie the perceptual and the introspective processes. More recently, Nichols and Stich (2003) defend a similar view of introspection for beliefs, calling the required cognitive

26 For detailed discussion of these features of Naïve Realism, see Martin (1997), (1998), and (2002).
mechanisms “monitoring mechanisms”; and they also suggest that this account generalizes to other mental states as well, as long as different “monitoring mechanisms” fulfill those other functions. I will label this model the *Monitoring Perceptual Model*.

Before moving on to the next version of the Perceptual Model, I would like to flag that I do not favor either the “scanner” or the “monitor” metaphors, insofar as they suggest that the relevant mechanism is constantly operating, which is an empirical matter and one I have no strong intuitions about. All I want to make room for is the possibility of mental state detection below the threshold of consciousness, but without that detection necessarily being constant (it may or may not be – I take no stand on that). Furthermore, Nichols and Stich speculate about the precise nature of mental state monitoring, which is yet another empirical matter that is far from fully investigated even by those who specialize in it. As such, I would like to make some distance between the simple Monitoring Perceptual Model, like Nichols’ and Stich’s version, and my preferred version of the Perceptual Model on this end of the perceptual spectrum. So a commitment to the Monitoring Perceptual Model is merely a commitment to us having fully conceptualized beliefs about one’s mental life, acquired contingently through some dedicated mechanism capable of operating below the threshold of consciousness.

Finally, there is a third version of the Perceptual Model, which is highly related to the Monitoring Perceptual Model (if not its variant) – a model I shall term an *Attention-Based Perceptual Model*. This is the view largely associated with Lycan, and particularly his (1996), though Goldman assign attention a significant role in his (2006). Lycan focuses on the apparent voluntary control we can exercise over our introspective access to certain mental states, including being able to focus on different parts of your visual
field, as if similar attention mechanism as in perception were involved. This observation leads him to think that we possess dedicated attention mechanisms, which he interchangeably terms “internal” or “introspective” (e.g. 1996, 16).

Although Gertler (SEP entry on “Self-Knowledge”, 2011 version) groups Lycan’s view together with Monitoring Perceptual Models, like Armstrong’s, it is best to distinguish Lycan’s view from simple monitoring view for our purposes. The monitoring view does not require a complex mechanism; simply adding “I believe that _____” to propositions in the “belief box” could qualify as a monitoring mechanism (which is, in fact, Nichols’ and Stich’s view). Lycan’s view, however, requires a different, or perhaps even a further, selection mechanism to choose which mental states (or contents, or properties) to focus on. It does appear at times that Lycan attributes both the monitoring and the selective powers to a single cognitive mechanism, but that need not be the case, and particularly not for a pluralist; and it is certainly not part of Armstrong’s or Nichols’ and Stich’s views. So, throughout this chapter, to be clear on which claim is at issue – whether it concerns simple monitoring or focused selection – I will keep these two versions of the Perceptual Model distinct, the Monitoring version assumed to be just that, without any added features, and the Attention-Based version encompassing only the selective aspects of introspection, without the monitoring aspects (which very well may be a departure from Lycan’s actual view, but that need not concern us here).

In order to deal with the skeptical worries outlined at the beginning of this section, and more importantly to accommodate all the data regarding our awareness of phenomenology we need to welcome all three versions of Perceptual Models just outlined. Let us consider the data first, and the epistemic desiderata second.
A Motivating Case

In the previous chapters, I used the example of phenomenology “hijacking” our attention to illustrate the need for a Perceptual Model in a complete account of introspection. A closer look at a case of “hijacking” can also reveal the need for multiple versions of the Perceptual Model, which are all now on the table, so let us examine such a case in greater detail.

Imagine going for a pleasant walk in the woods. The air is fresh, the scenery is impressive, and the temperature is comfortable. You pause to examine an unusual bird up above, when all of a sudden a sharp pain shoots through your leg. Surprised, you examine the place where the damage might be, but finding nothing out of the ordinary, you consider the feeling of pain itself, so that its particular characteristics might give you signs of an explanation. Deciding that it was just a muscle twitch, and not a sign of some serious trouble, you continue your walk.

The case of sudden pain is particularly striking and familiar, but a similar story can be told almost about any mental state. One could be equally surprised by one’s own desires, thoughts, and emotions. For example, one might find themselves wanting pumpkin soup, which they usually dislike; or catch themselves thinking about baseball, which they normally find boring; or discover themselves to be relieved about a dinner party being cancelled, even though they have been looking forward to it all week. In all these instances, the surprising nature of the mental state involuntarily makes us focus on it and its content in particular, to make sure that one is not mistaken about one’s experience.
The element of surprise is what is crucial about all these cases of mental states “hijacking” our attention, which illustrates the same two points as surprise about something one sees does. First, one is registering on some level what is happening before the surprising element actually occurs. This is illustrated by the fact that one returns to the surprising mental state in order to examine it further. And second, the surprising element grabs our attention and involuntarily redirects it.

Not all surprises have either these two elements, of course. Sometimes one is entirely unaware of what is going on before the surprise happens, as when someone jumps out in front of you from behind a door, screaming “Boo!” Some of the milder surprises, though, do involve prior registering of the relevant environment, and because of this prior registering does the relevant cognitive mechanism redirects one’s attention – without person-level awareness or control.

Likewise, shifts in attention, visual or cognitive, need not be involuntary; we are clearly able to direct our attention at will in many situations. We can examine a painting at length, methodically observing it from one corner to the other. Similarly, we can mentally go through a proof, provided it is not too long or complicated, checking every premise and inference one by one. In fact, it is this ability to shift one’s (cognitive) attention voluntarily, which provides the basis for Lycan’s view. And to that extent, it is clear that the Attention-Based Model of is a necessary part of a complete view of introspection.

However, since I am interested in providing a complete account of introspection, we must consider all cases of surprise, some of which are about our mental life and are best
illustrated by involuntary shifts in (cognitive) attention, as described above. Such cases are much more striking and significant for my purposes, for they illustrate a feature, which is much more closely associated with the Perceptual Model than with reasoning-based models of introspection – namely, the lack of constant control and accessibility to every step of the introspective process.

Thus, what the described cases of surprising mental states illustrate is not only that they are best accommodated by the Perceptual Model, but that at least two versions of the Perceptual Model are required. First, we need the resources of the Monitoring Perceptual Model to explain how we register our own mental states prior to the surprise, so that the surprise is even possible. And second, we need the Attention-Based Perceptual Model to explain how we can involuntary focus on a particular mental state and its content in particular – which, incidentally, we then go on to scrutinize entirely voluntarily, once again illustrating both ways of shifting one’s attention.27

As for the Constitutive Perceptual Model, one does not need a particular case to see the need for it, since the motivation to include it in a pluralist account of introspection are largely epistemic, to which we presently turn.

**Epistemic Considerations**

27 As mentioned earlier, it is possible to combine the Monitoring and the Attention-Based Models into a single mechanism that can perform the functions described by the two models (which, as I said, is more akin to Lycan’s view than the isolated Attention-Based Model is). However, in an attempt to keep the necessary features of introspection as clear as possible, I have kept the different, necessary functions of the Perceptual Model separate in its different versions. Combining these functions into a single mechanism does not detract from the fact that a multi-function Perceptual Model is required in a complete account of introspection.
The main motivation for the Constitutive Perceptual Model comes from an intuitive conception of experiences, particularly of sensations and perceptual ones. On this conception, there is a layer of experience to which we can have complete and infallible access, because that layer depends on us for its existence. This is particularly salient if we consider non-veridical perceptual experiences, i.e. illusions and hallucinations. If I am hallucinating a pink cat in the corner of a cat-free room, intuitively, whatever I “detect” about this hallucinatory cat, that is all there is to its existence. This is, in fact, what I take to be the motivation for the *seemings*- or *appearance*-talk, to signal that we are considered with a layer of experience to which only its subject is privy. In other words, there is a layer of our experiences, the reality of which *is* their appearance, and the intuitive view is that phenomenology of mental states is part of that layer.28

This view of experience has been criticized by Williamson in his (2000) “anti-luminosity” argument, in which he attempts to establish this conception of experience to be implausible. Such arguments, however, do not directly undermine the minimal infallibility claim I am proposing, for two reasons. First, Williamson assumes the layer of experience that can be infallibly accessed to be thicker than I am suggesting. For example, he argues that one does not know that one feels cold at a given time; but the belief that can result from infallible access I am proposing is a merely demonstrative one, partially constituted by the relevant phenomenology. As I will discuss shortly, the demonstrative belief could be infallibly formed without the subsequent belief that one feels cold being infallibly formed.

---

28 Although this conception of experience is usually associated with Descartes, similar ideas, though in more modest form, can be found in relatively recent work, e.g. Kripke (1980), Hill (1991), and Sturgeon (2000).
Nonetheless, there are other potential costs to views involving an infallibility claim, such as the set of worries associated with the “veil of perception”, which I address in outline later in this section. However, denying this intuitive conception comes with its own costs, which is the possibility of error where intuitively there is none. The debate about the nature of experience is rich and a deeply divided one, and I shall not digress into it here. I will just flag that for those who are drawn to the intuitive conception of experience as in the case of a pink cat, the Constitutive Perceptual Model is an undeniable part of a complete pluralist account of introspection, since it is necessary to explain our access to hallucinations, and also to demonstrative thoughts and beliefs, the contents of which are not fully conceptualized. And I, for one, find this view plausible. However, even one is unmoved by the pink cat case, and so does not feel the need to accept the Constitutive Perceptual Model, I have already argued for two other versions of the Perceptual Model in the previous section, so the pluralist account so far already contains other versions of the Perceptual Model. To reiterate a point from the beginning of this chapter, part of my goal is simply to demonstrate that pluralism is compatible with different kinds of conception of experience and phenomenology, and not to present a knockdown argument for any one of them. Thus, the Constitutive Perceptual Model is necessary if one finds plausible the conception of experience on which phenomenology can be (though need not always be) infallibly accessed; otherwise, a combination of the Monitoring and Attention-Based Models will suffice. Either way, the Perceptual Model is necessary for a complete view of introspection.

Consider now how combining different versions of the Perceptual Model can create a more plausible, complete view of introspection. The main problems with the Constitutive
Perceptual Model, as mentioned earlier, is that on the one hand, it leads to an infallibility claim, which can become implausible very quickly if it applies to a significant enough portion of our mental life. But on the other hand, it can also lead to beliefs that cannot be integrated with the rest. To address both of these problems, it is crucial for a pluralist to rely on a combination of Constitutive Perceptual Model together with the Monitoring and Attention-Based Perceptual Models. The key is to rely on the Constitutive Perceptual Model only to grant us access to the phenomenology and other non-conceptual parts of experience, if any, so that we may form a demonstrative belief about it. At this point, we are free to fully conceptualize that belief, and to integrate it with other beliefs, for example, through reasoning. This simple two-step process is part of the Monitoring Perceptual Model, and one that explains how the fleeting beliefs partly constituted by our phenomenology can be integrated into the larger belief system. While these fleeting beliefs exist, we conceptualize their contents – and the existence of the conceptualized result need not be as fleeting as the unconceptualized phenomenology itself. In essence, my purpose for introducing the Monitoring Perceptual Model in a pluralist account is to use it as a vehicle of transforming, sustaining and integrating our infallibly acquired phenomenological beliefs, resolving the issues of the Constitutive Perceptual Model. And naturally, we can then go on to focus on particular aspects of our mental states, either voluntarily or not, with the help of the Attention-Based Perceptual Model.

The combination of these three versions of Perceptual Models, though the Constitutive and the Monitoring ones in particular, also explains how we can infallibly access our phenomenology while leaving room for error in our access to other mental states. While the fleeting phenomenology-containing beliefs are indeed infallibly
acquired, the process of conceptualizing their contents is fallible. The difference in these two processes not only allows for error, but also introduces the possibility of error in the right place. Consider describing your current visual experience to a blind person – no matter how detailed your description is, it would still feel like you are cheating by using color words, even if the blind person claims to understand their referents. The same difficulty arises if you try to describe the suffering from pain to a pain asymbolia patient.

These examples illustrate three points. First, our phenomenology is not (fully) conceptualized when we first experience it, and there does not seem to be any room for an appearance/reality distinction, when it comes to apprehending the phenomenology of one’s experience. Second, we are capable of somewhat successful conceptualization of our phenomenology while we are still experiencing it, and we are able to recognize that the subsequent fully conceptualized belief can be compared with the “raw” phenomenology we started with, even if temporarily. And third, something of our “raw” phenomenology is frequently lost during conceptualization, particularly when the experience is complex. This shows that the proposed combination of the three Perceptual Models divides our awareness of our experiences into error-free and error-prone in exactly the right place.

**The Possibility of “Effortless” Introspection**

Before examining further objections I have not yet considered to the three versions of the Perceptual Model outlined in this chapter, it is worth highlighting a particularly distinctive feature all three versions of it share – and perhaps all Perceptual Models do – which will prove to be particularly important in the next chapter. One way of putting the difference between the Perceptual Model and the other two models we have examined,
the Third Person Model and the Transparency Model, is that the Perceptual Model feels effortless while the Third Person Model and the Transparency Models are effortful. The latter, economical models essentially offer *methods* that one needs to consciously apply in order to introspect one’s mental states. By contrast, whatever cognitive work is required by the Perceptual Model, most of it happens below the threshold of consciousness, and so does not feel effortful. I am careful to say “feels effortless” (or “effortful”) because it does not make sense to describe the Perceptual Model as *being* effortless, since clearly cognitive resources are being spent during introspection, even if they are below the threshold of consciousness. To see the difference, compare the Perceptual Model with Shoemaker’s Constitutivism, according to which introspection is not even a process but a structural feature; that is the only genuinely effortless view. This is not an entirely terminological point, for Schwitzgebel (e.g. in his 2010 *SEP* entry on “Introspection”) does not think that introspection can be effortless at all, and so it is important to carefully describe the level of effort the Perceptual Models involve.

There is more to this contrast between the Perceptual Model feeling effortless and the Third Person Model and the Transparency Model feeling effortful. Instead of requiring conscious effort to *begin* the introspective process, as the economical models do, the Perceptual Model allows for situations in which the demand for our person-level attention happens only at the *end*. As already described, instead of leaving the choice of when and how to introspect, the Perceptual Model allows for, though does not require, the feeling of having one’s attention “hijacked” to its results, just as seeing strange things in the periphery of our vision makes us turn our heads and examine them directly, until we are satisfied with our analysis of what we saw (or seemed to see). This feature of the
Perceptual Model will become very important in the next chapter. But for now, this discussion only serves to keep the Perceptual Model from being ruled as an account of introspection, as Schwitzgebel would have it. I defer a more thorough discussion of the effortlessness issue to the next chapter.

**Objections to the Perceptual Model**

Turning now to the few remaining objections to my preferred versions of the Perceptual Model, let us consider first a worry concerning the infallibility of our introspective access that the Constitutive Perceptual Model offers. One might worry that any infallibility claim concerning our access to the mental might be too strong. The best response to this objection is just to emphasize the very limited nature of this account, in that it grants access only to our current phenomenology. It should not be so implausible that current phenomenology can be accessed perfectly; in fact, it is in line with the original take on the concept of phenomenology as “what it’s like” to experience something, which suggests infallible access, as with all appearances. And one can certainly go wrong in the process of conceptualizing the accessed phenomenology, so this account does not guarantee the truth of non-demonstrative, conceptualized beliefs about our phenomenology. In short, it seems to be a positive feature of the account that it can give us infallible access to current phenomenology, rather than a problem.

The other objections to the Constitutive Perceptual Model become irrelevant when the account is couched within a pluralist framework. First, one might object that it does not give us enough self-knowledge (e.g. Wright 1989). If we accept that only our access to unconceptualized phenomenology is infallible, and that its conceptualization could
introduce error into our beliefs about it, then we might not (ever) have beliefs about our phenomenology that could count towards our stock of self-knowledge. There are ways to resist this objection, but a pluralist about introspection can simply bite the bullet, because she has at her disposal other ways to gain self-knowledge, which may be less error-prone, such as the Transparency Model. Similarly, a pluralist need not be concerned that the Constitutive Perceptual Model cannot give us access to the contents of our beliefs and other propositional attitudes (objection from Boghossian (1989)), because there are other models of introspection, which grant us that access.

As for the Monitoring and the Attention-Based Perceptual Models, the objections to them target only the extent to which they make introspection in any way privileged or special. None of the objections bring into question the proposed mechanism itself – understandably so, since the precise nature of such mechanisms is an entirely empirical question. And as I argued in the previous chapters, it is best to leave the question of privilege and specialness of introspection until its mechanisms have been chosen and defended.

Now that all the necessary models of introspection have been examined and defended against objections, it is time to take stock.

**The Perceptual Model in a Pluralist Framework**

One common theme in all the objections addressed across the different models of introspection is that each model individually has been found inadequate in its range of application. This theme highlights something very important about introspection, which is that given the enormous variety of our mental states/events/processes, it is
unreasonable to expect a single mechanism, whether dedicated or not, to access all of them with a tolerable degree of accuracy. And yet we seem to manage it (skepticism about introspection aside).

Though it is true that the burden to explain this fact is on every account of introspection, it presents two special problems for any version of the Perceptual Model. The first problem stems from the severe limits on its range that a Perceptual Model has to suffer. Take the Monitoring Perceptual Model as an example – it only explains our access to our conceptualized phenomenology, nothing more. But what if we wanted to explain our access to other mental properties with a Perceptual Model? The problem is that physical space in our brains is quite limited, and there just might not be enough of it, should we insist on a dedicated introspective mechanism for every aspect of our mental lives.

The second, related problem is that, given that dedicated cognitive mechanisms occupy brain space, and that nothing occupies brain space without being evolutionary advantageous, the burden is on the pluralist to show that there is an evolutionary advantage to having specifically a perception-like, dedicated mechanisms for introspection of any sort. This evolutionary concern becomes the more pressing the more distinctive dedicated mechanisms are required. These two concerns develop precisely because the Perceptual Model is placed in a pluralist framework, and as such they present the final obstacle to the pluralist account I have been arguing for.

Although the pluralist framework is the root of the outlined concerns, it is also the source of the answer to them. Certainly we do have cognitive space for the different
perceptual faculties – sight, hearing, smell, touch, and taste – so the very idea of having similar mechanisms that nonetheless specialize in different inputs should not be problematic. What would be problematic is if I proposed distinct and dedicated introspective mechanisms for every mental states type. But that is far from my suggestion. And since my proposal is quite modest in terms of required cognitive space, only the purely empirical matter remains of discovering where these mechanisms are located and how they operate, assuming we in fact possess them.

As for the evolutionary concern, there are two things to say in response. One is that it is indeed difficult to say why detecting phenomenology in particular would be useful in evolutionary terms; why it would be useful to detect other mental features is more thoroughly discussed, e.g. by Shoemaker. My own, currently underdeveloped, outline of the explanation is that phenomenology is very useful in directing our attention to the experience to which it belongs. Phenomenology allows us to judge experiences to be pleasant or unpleasant – i.e. helps us determine the valence of an experience – and ultimately determine the appropriate actions in response to those experiences. In short, phenomenology is a crucial factor in shaping our motivation to act. As such, it is useful to be aware of it to make our own motivations understandable, which is in turn useful in deciding the most appropriate course of action – for not all of our impulses are worth acting on.

If this evolutionary account is unsatisfactory, consider the possibility that our ability to access our own phenomenology could have been a side effect of something else developing that is clearly evolutionarily useful. Moreover, no matter why, on the evolutionary scale, we developed the ability to introspect, and detect phenomenology in
particular, the fact remains that we do have access to a very wide range of mental states/events/processes. And if the Perceptual Model is the only one that can explain our access to some of those states, then we should be asking ourselves whether we are more comfortable with denying that we have access to those states at all than with the Perceptual Model.

There is one final worry about the Perceptual Model in a pluralist framework worth considering. When first introducing the Perceptual Model, I claimed that it can be used to account for our access to any mental state without the help from any other introspective account; and throughout the discussion of the potential flaws of the Perceptual Model, very little has come out to apparently hedge that initial claim (and perhaps only the concerns expressed in this section go some way towards that). So one might conclude that the Perceptual Model is actually incompatible with a pluralist account, since it is powerful enough to provide a complete account of introspection all by itself (setting aside whether this would be an independently plausible account of introspection).

The appropriate response to this worry illustrates once again something important about my ultimate goal, which is to provide an accurate and complete account of introspection, rather than a theoretically and minimally necessary one. As long as completeness is part of the goal, we cannot ignore the fact that we do use the Transparency and Third Person methods in real life quite frequently: we rely on the external world to help us access our beliefs, and we use our behavior as a guide to our mental life as well. So even if the Perceptual Model is theoretically powerful enough to provide a monist, unified account of introspection, it does not mean that such an account
would accurately describe *all* the ways in which we can access our mental states. In short, the fact that the Perceptual Model easily generalizes does not make the economical models, and with them the whole pluralist framework, unnecessary.

**Conclusion**

This concludes my discussion of the existing models of introspection and the master argument for pluralism, a view consisting in the Third Person Model, the Transparency Model, and three versions of the Perceptual Model. The essence of the argument for pluralism has been that our mental life is just too diverse to expect a single model of introspection to explain how we can access as much of it as we apparently do. And in order to provide a complete account of introspection we have to rely not only on the economical models like the Third Person Model and the Transparency Model, but also seek help from the extravagant Perceptual Model. The precise restrictions on each of the models of introspection, placed by the differences between mental state types as well as the objections to each of the models, have formed the topic of this chapter.

This chapter also concludes the argument from limitations of the existing models, showing the need for pluralism about introspection. In the next chapter, I offer a different argument for pluralism, one that is based in a case that requires several introspective steps, and a different model to accommodate each one.
Chapter 5: A Complex Case of Introspection

In the previous chapters, I argued that we cannot account for our access to the entire range of mental states with a single model of introspection. That is, no single model of introspection is sufficient to explain our access to all mental states. So a complete account of introspection has to consist of several different models, sometimes seen as competing against each other. I proposed and defended a combination of the Third Person Model, the Transparency Model, and the Perceptual Model as minimally required for a complete account of introspection.

In addition to the limitations of each of the considered models raised in the previous chapters, there are also example-driven considerations that support pluralism about introspection. In particular, the more complex, multi-step cases of introspection naturally invite a pluralist account. One of these cases is the focus of this chapter. After describing the case in detail, I show that pluralism about introspection is required to provide the most natural account of the case. And since the case under consideration is easily generalizable, realistic, and commonplace – not just a thought-experiment – we should conclude that pluralism is required for a complete and realistic account of introspection.

This chapter is divided into six substantive sections. In the first two sections, I present the motivating case and analyze its structure, to better see its generality. The three subsequent sections are dedicated to the argument for pluralism based on the analyzed case. Crucially, this argument hinges on establishing the necessity of the Perceptual Model to accommodate the central case, since it is the only non-economical
model of introspection considered here. And in the final section I address a worry about pluralism based on the intuitions about the unity of the mental and of introspection.

The Case

Consider the following scenario: You meet a good friend of yours on the street, and engage in a casual conversation. Your friend makes a joke at your expense – nothing unusual or particularly rude – and yet you find yourself moved to make a rude comment in return. But you restrain yourself, noticing that your rudeness would have been greater than the harmless joke justified. As a result of this observation, you realize that you must have been angry about something else in the first place, and not about the joke. And upon further reflection about the earlier parts of your day, you pin-point the real object of your anger: rather than being angry at your friend, you are angry at yourself for oversleeping that morning and being late to an important meeting. With this new and better understanding of your background emotional state, you have an easier time taking the joke without holding a grudge against your friend, and so carrying on with the conversation.

Even if one does not find themselves in this particular situation very often, this case does have a pattern, which, once extracted, can be clearly detected in many daily situations. In essence, this case exemplifies learning something about yourself from unactualized behavior – specifically, learning something other than that one did not actualize some specific behavior. Denying that this happens quite frequently would be extremely implausible. Even so, the precise structure of this example is not entirely obvious, since the case is clearly complex. But for my purposes, only three steps matter
in the basic structure of this example. That is because each of the three steps involves learning something new about one’s own mental states, and more importantly, each of those steps most naturally accommodated by a different introspective model – the Perceptual Model in the first step, the Third Person Model in the second step, and the Transparency Model in the third step. The rest of the chapter is dedicated to showing this to be the case.

Before moving on to the three steps though, I should note that throughout the discussion of the case I use success terms like “discover,” “realize,” and “learn”, to name a few, all of which suggest that the resulting state is knowledge-like. However, I would like to avoid this inference and the discussion of whether introspection of any sort yields knowledge – for now. Rather, I have chosen these terms because they also have a significant amount of phenomenological accuracy, in that they capture very well how it feels for the subject in the described case. And to the extent that it feels like we know what we self-ascribe throughout this case, the success terms are entirely appropriate. With this caveat in mind, let us now examine the three steps in detail.

The Analysis of the Case

Turning now to the case described in the previous section, the first significant step consists in one’s noticing the motivational tug, such as an intention or desire to behave in a certain way – in the given case, to make a rude comment towards your friend. Typically, only the potential behaviors that somehow fail to be in line with the current situation will get so noticed, but it is not inconceivable that a generally more self-aware person might notice some of the ordinary motivational states as well. It is, however,
more natural to assume that the desired/intended behavior is in some sense currently inappropriate. Moreover, what is important at this stage is that the said behavior remains unactualized, i.e. one has restrained themselves from making a rude comment. This makes it clear what is being noticed – the motivation to behave in a certain way, rather than actual behavior – and it is important to keep these separate.

After noticing the motivation to behave, and restraining oneself, comes the second step, which is realizing something about one’s mental states. There are three points worth noting about this stage. First, it tends to involve or be accompanied by some degree of surprise; in the case as I have told it, it might be expressed as the (internal) reaction, “Wow, I’m actually angry!”

Second, and more importantly, the content of this surprising realization is not yet as precise as it could be. That is, what one realizes at this stage is simply that one is angry, as opposed to one is angry about oversleeping on the one day this week when one really needed to be perfectly awake and on time for an important meeting. The latter realization involves details about the precise content of a mental states, or in this case its precise object – and that sort of information is not available at the second stage. In short, what one learns in the second step is just that one is in some blanket mental state, without learning much about the content or the object of this mental state.

The third and the most important feature of the (not fully specified) realization in the second step is that you learn that you are in a mental state other than the motivation to act noticed in the first step. This is precisely why this step is differentiated from the first, noticing step – because you learn something new about your mental states, on which the
first step was silent. Moreover, what one learns in the second step cannot be assimilated to the first step by describing the case as noticing (in the first step) both that you had a certain motivation and that you were angry at the time. The reason for this is that learning in the second step that, in our case, one is angry requires having noticed that one had a certain motivation in the first place. The realization cannot occur without prior noticing that one has a motivation to act in a certain way. The noticing sparks the realization, and so they cannot be part of one and the same step. This is also true of the second and third steps – that each of them depends on the immediately preceding step – which is why they are treated as separate.

Moving on now to the third step, one adds to the discovery made in the second step by accessing one’s beliefs about its object, which would eliminate the surprise of the mentioned discovery. In our example, on further reflection, made in the third step, the realization that you are angry is supplemented by the specifics of that mental state – that you are angry at yourself about oversleeping and being late for a meeting, and not about your friend’s joke. The contrast between this step and the previous one is worth emphasizing: While the initial discovery that you are angry may be mildly surprising at first, the whole point of focusing on that discovery is to eliminate the surprise and explain to yourself why you are angry. And that is the purpose of the third step. In other words, the latter two steps have different accompanying phenomenology – one surprising, the other illuminating – which gives us prima facie reason to keep the two steps separate.

It should now be clear why I have chosen to highlight the three steps of the case that I did – because each one involves self-ascribing a different mental property each time. In the first step, we self-ascribe the motivation to behave in a certain way; in the
second step, we self-ascribe a broad mental state, like anger; and in the third step, we form a belief about the object of one’s anger, which in the case of anger also happens to be its specific intentional content and its root cause.

With the details of the case now on the table, let us now consider the extent to which the given case supports pluralism about introspection. For we are far from establishing pluralism simply by pointing out that a single case can involve a three-step introspective process. By itself, this shows nothing about the mechanisms underlying each of the steps. However, each of these steps intuitively exhibits certain features, which do bear on the underlying introspective mechanisms.

The argument for pluralism in this chapter involves matching different types of mechanisms to different steps of the central case. First, I identify a general feature of mental processes – a type of onset they might have – and make a contrast between the three steps of the described case based on this feature. In particular, I argue that only the first step possesses a certain type of onset, while the remaining two steps exhibit a different kind of onset. Second, I argue that only one model of introspection, namely, the Perceptual Model (an extravagant model) could possess the right type of onset that accurately describes the first step; the other two models of introspection, the Third Person Model and the Transparency Model (economical models), could only accurately describe the remaining two steps. And third, I present and resolve a difficulty in what should appear to be a straightforward application of Leibniz’s Law. The result of this three-part argument is that at least two models are shown to be required to accommodate the case that drives this chapter – and importantly, one of those models has to be extravagant.
while the one other(s) can be economical – thereby establishing pluralism about introspection.

During the third and final part of the argument, I go on to make an additional argument for needing both economical models, the Third Person Model and the Transparency Model, to do justice to the complexity of the central case. This additional argument is not strictly necessary to establish pluralism – merely showing the need for both an extravagant model and an economic model is enough – but it does help in establishing the particular type of pluralism that I think is minimally necessary to account for our access to the entire range of our mental states.

*The Argument for Pluralism, part I: Active v. Passive Onset*

Consider a familiar example of the “cocktail party effect” – you are in a large crowded room, filled with conversations among small groups of people. You are deep in conversation with someone, and although you might be registering the generally high level of noise, you are too engrossed in your own conversation to notice what the people outside of your group are saying. And yet, when your name comes up in a conversation across the room, you can all of a sudden hear it mentioned despite all the noise, without made any apparent effort to do so; in fact, you have been trying not to hear what the people outside of your group were saying.

The relevant contrast in this scenario is between actively engaging in a conversation, part of which is actively listening to your interlocutor, and finding yourself hearing – or overhearing – your own name across the room. Hearing, or overhearing, is something that you cannot help but do – at first approximation this aspect of hearing
might be described as constant, or effortless, or automatic. Whichever term seems most appropriate, the relevant feature is whatever makes hearing your own name interfere with your own conversation. By contrast, listening to your interlocutor in a noisy room requires effort on your part. In fact, if the topic of conversation is sufficiently interesting or difficult, this effort would be required even if you were in a quiet room, though of course it may increase dramatically in less than ideal conditions.

There are, of course, other effortful processes, besides being engaged in an interesting conversation. And sometimes it is proposed that introspection must be so effortful. This prima facie inviting thought has been turned into one of six conditions, or requirements, on any account of introspection by Eric Schwitzgebel:

*The effort condition*: Introspection is not constant, effortless, and automatic. We are not every minute of the day introspecting. Introspection involves some sort of special reflection on one's own mental life that differs from the ordinary un-self-reflective flow of thought and action. The mind may monitor itself regularly and constantly without requiring any special act of reflection by the thinker—for example, at a non-conscious level certain parts of the brain or certain functional systems may monitor the goings-on of other parts of the brain and other functional systems, and this monitoring may meet all five conditions above—but this sort of thing is not what philosophers generally have in mind when they talk of introspection. (Schwitzgebel, SEP entry “Introspection” 2010)

There is an important distinction to be made along this line, as I have attempted to illustrate with the “cocktail party effect” case. Exactly where the contrast lies, though, is
somewhat difficult to tell. For every model of introspection is likely to require at least some cognitive work to be completed below the threshold of awareness, maybe even necessarily so. For example, reasoning and particularly any perceptual mechanism require pre-conscious cognitive work. In fact, it would be surprising if any conscious result did not require pre-conscious work. But just because cognitive work is pre-conscious does not mean it is entirely effortless (or automatic); and it is very tempting to interpret the term “effortful” to mean something like “requiring conscious effort,” for it is hard to gauge one’s own efforts below the threshold of consciousness.

So instead of using the contrast between effortful and effortless processes, I will appeal to a distinction between active and passive onset of the relevant processes. That is, a process that must be started through conscious, or person-level, voluntary effort is an active onset process. Engaging in an interesting conversation, or drafting an article, for example, are like that. By contrast, passive onset processes may begin without such effort. Importantly, the lack of conscious effort is not required, but merely possible, for passive onset processes. Hearing and other perceptual processes are like that – they are plenty effortful, but neither consciously nor at the outset. Only their results tend to be conscious (and even so, not always). Since passive onset processes are also capable of an active process, the crucial contrast between types of mechanisms discussed in this chapter will be between mechanisms that require an active onset that those that do not (i.e. those that are capable of a passive onset). I submit that some introspective processes require an active onset and some do not, and that there is a mix of such processes present in the case central to this chapter.
But perhaps the very idea that an introspective process should be capable of a passive onset might seem unusual, particularly if one is sympathetic to something like Schwitzgebel’s requirement that introspection should be effortful in some relevant sense. However, I see no reason to rule out by fiat passive onset processes from being introspective. The radical differences between views on introspection defended in the literature, and examined throughout this dissertation, shows that conceptual analysis or even an appeal to intuitions cannot reveal the kinds of mechanisms that may be involved during introspection. So I do not see how one can justify a necessary condition on all introspective mechanisms of the sort Schwitzgebel proposes. At best, the effortfulness of introspection – or as I have been describing it, having an active onset – can be only an expectation, and not even clearly a desideratum (for I see no benefit in having an active onset at such), of what the final analysis of introspection might be like. But I have already argued, in the introduction, for a generally more inclusive approach to the term introspection.

Turning now to the case that is the focus of this chapter, recall that it has three pivotal points. First, after talking with your friend and hearing the joke at your expense, you notice the desire and/or intention to make a rude comment in return. Second, you use this observation to discover that you are angry. And third, you determine the object of your anger.

We are now in a position to talk about the crucial contrast between the first step and the rest. The contrast is simply this – the first step has a passive onset while the other two steps have an active onset. The way the case is told, you are in the middle of a casual conversation with a friend, and you are not also wondering about your own mental states.
And yet, you **still** manage to notice that you are about to make an angry comment. This strongly suggests that coming to pay attention to the relevant motivation to act is made possible by a pre-conscious, involuntary process. Thus the first step requires an introspective process capable of a passive onset, or put differently, one that does not require an active onset.

By contrast, once you notice the motivation to make a rude comment, your (cognitive, conscious) attention is directed to it, which is all it takes to make the discovery of your anger have active onset. The noticing itself delivers the conscious result that acts as the active onset of the process, which leads to the discovery of your anger in the second step. Similarly, the discovery of anger then takes up your (cognitive, conscious) attention, which is the active onset of the third introspective process – the one resulting in your belief concerning the object of your anger. Thus the second and third steps involve an active onset; and I as will argue in this chapter, these two steps require an active onset, i.e. they cannot be explained in terms of a mechanism that is capable of a passive onset.

Let us now examine the three models of introspection I have been defending – the Perceptual Model, the Third Person Model, and the Transparency Model – to determine which kind of onsets each of them requires.

*The Argument for Pluralism, part II: The Three Models of Introspection*

It is easy enough to see that the Perceptual Model does not require an active onset, and can accommodate a passive onset. In fact, this model is prototypically “effortless,” and the very model of introspection Schwitzgebel is apparently intent on ruling out by
making the effort condition a necessary one on any account of introspection. For just as we might choose to (try to) attend to a mental feature – as we can do with perception, by directing our visual attention – we can simply notice something about our mental states (same as with perception of our environment), which does not require person-level effort on our parts. This is particular true if we consider the Monitoring version of the Perceptual Model, which, as its name suggests, is meant to monitor our states sometimes below the threshold of conscious awareness. In this respect, at least, the Perceptual Model is very much like perception – a parallel that is otherwise quite tenuous (and which need not concern us here, since this worry has already been addressed in a previous chapter dedicated to the Perceptual Model).

However, it is not as easy to see why the Third Person Model and the Transparency Model require active onsets, in that they require directing one’s (cognitive) attention to the relevant question or topic in order to deliver the relevant second-order belief. Importantly, the claim is not that the Third Person Model and the Transparency Model metaphysically or conceptually require an active onset, but rather that for various psychological reasons these models are only useful if, for the most part, the mechanisms they instantiate have active onsets. These observations are not immediately obvious, so allow me to expand, starting with the Third Person Model.

As the name suggests, the basic idea behind the Third Person Model is that however one manages to attribute mental states to other people, that same method can surely be extended to partially explain mental state self-attribution. And broadly, the way we attribute mental states to other people is through observing their behavior, which then feeds into an inference to the conclusion about other people’s mental states. But
extending this account to one’s own mental states is not as simple as noticing that this account is, in some sense, subject-insensitive: that it does not matter whose behavior you observe and then attribute mental states to, thus allowing this model to explain one’s own mental state attribution. For there is some subject-sensitivity to this account, namely the fact that one has to observe a person’s behavior. While we can literally observe other people’s behavior, through our senses, it does not seem that in most everyday situations we have the same type of (sensory/perceptual) access to our own behavior.

In fact, recognizing one’s own behavior is a significantly different task from recognizing someone else’s behavior. For example, getting back to the complex case that motivates this chapter, before I can infer that I am angry in the second step, I need to realize that I was about to be rude. By contrast, no conscious realization about behavior is necessary when someone else’s mental states are at stake. For example, I do not need to take conscious note of someone else’s rudeness before I attribute to them the state of being angry. Frequently, I might not even consciously recognize someone else’s behavior as rude before I say (to myself), “Wow, she’s angry!” However, the same does not seem to apply when one’s own behavior and mental states are concerned – you really do need to (consciously) notice your own rudeness before inferring that you are angry.

One might object that there may be sub-consciously operated mechanisms that track my dispositions to behave and can thereby provide sub-conscious inputs for the Third Person Model. In fact, such a view should be particularly appealing to someone who finds tracking of behavior to be a source of self-knowledge, as I have suggested. Indeed, I find this view of tracking behavioral dispositions entirely plausible, but it does not pose a problem for a pluralist. The sub-conscious tracking of behavioral dispositions
is nothing other than a monitoring mechanism, as described by the Monitoring Perceptual Model. This proposal, then, is essentially to supplement the function of the mechanisms underlying the Third Person Model with those of the Monitoring Perceptual Model, which supports rather than undermines pluralism. Moreover, adding a monitoring mechanism to the Third Person Model shows nothing about whether the mechanisms underlying the Third Person Model will thereby acquire the ability to operate passively. So the observation that the Third Person Model involves consciously attending to one’s dispositions (unless supplemented by the Monitoring Perceptual Model) remains intact.

In fact, this crucial observation – that noticing one’s behavior requires person-level attention – is frequently appealed to by psychologists in explaining the various biases, errors, and asymmetries to which we are prone. And it is this observation, which also explains why the Third Person Model, conceived of as a model of introspection, typically requires an active onset, rather than a passive one. As I mentioned earlier, this is not meant to be a conceptual restriction on the Third Person Model of introspection, but rather an a posteriori rule of thumb; after all, perhaps watching yourself in the mirror or seeing a video tape of yourself do something would allow for Third Personal mental state attribution to have a passive onset. Nonetheless, these unusual situations aside, it seems a basic fact about our own bodies not being so easily perceptually accessible to ourselves explains why the Third Person Model of introspection requires an active onset.

Something similar can be said about the Transparency Model. For example, in order for me to realize that I believe myself to be angry about oversleeping this morning,

---

29 See, for example, introductory psychology texts such as Aronson’s et al Social Psychology (2004) and Huffman’s Psychology in Action (2005).
I need to consider the previous events. Once again, the introspective process cannot begin without conscious effort. One needs to consider what else happened earlier that day, and form a belief about the real object of the anger; this is nicely accommodated by the Transparency Model’s claim that in certain cases introspection requires directing one’s cognitive attention “outward” (just like in Evans’ case of considering the possibility of third world war). Thus the Transparency Model also requires an active onset.

One might object that this conclusion about the Transparency Model as follows. According to this model, upon considering the question of my anger I immediately form a judgment regarding its object. Moreover, we frequently make judgments below the threshold of awareness, so it is possible for the judgment about the object of my anger to be formed without conscious effort, thereby eliminating the need for an active onset. However, this objection overgeneralizes some facts about judgments and reasoning. Although it is true that as a rule, reasoning and judgment formation can happen below the threshold of conscious awareness, this is not a fact that clearly extends to the reasoning and judgment formation involved in the Transparency Model. The reason for this is that such a picture would lead to us having a vast number of introspective beliefs, to match the vast number of pre-conscious judgments we are capable and do make. But we are clearly not overwhelmed by introspective beliefs about our mental states, so it is not advisable for a defender of the Transparency Model to exploit the possibility of pre-conscious reasoning and judgment formation. The sort of reasoning described by the Transparency Model does not perfectly fit the mold of general reasoning, as already discussed in the chapter dedicated to this model.
One might object that the previous line of argument oversimplifies the nature of subconscious reasoning. Presumably we do not use every belief in subconscious reasoning of a general sort (like modus ponens); rather, we use some beliefs but not others, and however it is determined which beliefs take part in subconscious reasoning, presumably the same can be applied to rules like BEL, thus preventing problematic overgeneralization described in the previous paragraph. In response, it is worth emphasizing the uniqueness of BEL, and any other potential transparent rule of introspection. In order to use BEL, the relevant question must first come up and we must consider our views on the matter. In other words, we must deliberate on the given topic. And this process of selecting the relevant belief to use in reasoning is hardly generic, or common to all reasoning. So I am doubtful that whatever can be said of reasoning in general is clearly applicable of the reasoning involved in the Transparency Model.

So far I have argued that neither the Third Person Model nor the Transparency Model can accommodate a passive onset, since it is unlikely that a sub-conscious judgment that $p$ (in the case of the Transparency Model) or a sub-conscious observation of one’s own behavior (in the case of the Third Person Model) can lead to a conscious belief about one’s mental states. But there is one more way in which these two models might turn out to have passive onsets. Nothing I have said so far rules out the following alternative picture of how the Transparency and the Third Person Models function: Suppose we start with a sub-conscious judgment that $p$ (in the case of the Transparency Model) or a sub-conscious observation of one’s own behavior (in the case of the Third Person Model); I did try to motivate the unlikelihood of the latter option, but I have not shown it to be impossible. We then make the appropriate inference at the sub-personal
level, as specified by each of the two models, to the relevant belief about one’s own mental states, \textit{which at that point remains at the sub-conscious level}. And only after the second-order belief is formed at the sub-conscious level, the appropriate cognitive mechanism, say cognitive attention, then may raise it to person-level consciousness. As long as a sub-conscious judgment or observation of behavior does not lead \textit{directly} to a conscious belief about one’s mental states, but first rests at the sub-conscious level, both models can function with a passive onset, rather than an active one.

The motivating idea behind this alternative structuring of the Transparency Model and the Third Person Model outlined in the preceding paragraph is that we have no way of knowing whether we have any beliefs about our own mental states at the sub-conscious level, and if we do, how many. I have so far tacitly assumed that all second-order beliefs must, at some point, become conscious, and frequently at the outset of coming to form a belief about our own mental states, but I have not defended this basic assumption. In fact, this assumption might be objectionable if one thinks that there must be a strong parallel between self-knowledge and knowledge of the external world. For we form plenty of beliefs about the external world at the sub-conscious level, which may remain there for any amount of time, or may be raised to consciousness via the appropriate mechanism. For example, many of our perceptual beliefs, like that there is a tree in front of me, might never become conscious, without sacrificing any of our ability to appropriately interact with our environment – for example, walking around the tree, rather than directly into it. So it might be the case with self-knowledge: we form plenty of beliefs about our own mental states at the sub-conscious level, and those beliefs remaining at the sub-conscious level does not clearly undermine our ability to reason –
reasoning and other rational activities being frequently cited as being possible in virtue of self-knowledge (e.g. Shoemaker).

This is a deep worry concerning some of the most fundamental questions about introspection and consciousness, and answering it fully is beyond the scope of this chapter. But I will make two replies: one is an outline of a full response, and the other is to show that even this alternative worrying picture cannot avoid pluralism, which is, after all, the primary goal of this chapter.

The full response that gets to the heart of the matter, which is a question about just how conscious introspection is, relies on our intuitions about self-awareness. Self-awareness, which is what introspection is meant to explain, does not seem to be exactly like awareness of the external world after all. For even if a person is, in some sense, aware of a tree in virtue of having a sub-conscious belief about the tree in front of them, which allows them to walk around it rather than directly into it, it seems intuitively less plausible that one is similarly self-aware if all that person has is a sub-conscious belief about their own mental states. The point is not meant to be a linguistic one, but rather about our intuitions concerning the different expectations of external world awareness and self-awareness. There seems to be something zombie-like about a person who might not have any conscious beliefs about their own mental states, and likewise there is something very unsatisfactory about calling such a person “self-aware.” Perhaps we attach a kind of self-understanding to a person by describing them as self-aware, which may require beliefs about one’s own mental states to be consciously available – the kind of understanding that may be absent when it is exactly parallel to awareness of the external world. What that understanding is, and whether self-awareness really must be
conscious, at least sometimes, are the difficult questions I have to set aside. But I hope enough of the intuition regarding the link between consciousness and introspection has been brought to the surface that the vaguely skeptical worry about most (or all) of our beliefs about our own mental life being sub-conscious can be comfortably ignored for now.

The other response to the sub-conscious introspection worry brings us back to the issue of pluralism about introspection. Recall that the troublesome alternative picture of how the Transparency and the Third Person Models function involves a sub-conscious inference from one sub-conscious state – either the judgment that \( p \) or the observation of one’s own behavior – to another sub-conscious state, the appropriate belief about the relevant mental state; and then this belief may become conscious via the general-purpose mechanism that makes the rest of our beliefs conscious. Although this picture does not require either model to have an active onset, contrary to what I have argued, this picture cannot escape pluralism either. For there are at least two distinct mechanisms required to complete the alternative introspective process: the reasoning mechanism at the sub-conscious level, and the “raising to consciousness” mechanism. Note too that the second mechanism, which raises the relevant second-order belief to consciousness, is nothing but a monitoring mechanism, and more specifically the mechanism of attention, which forms the basis for a type of Perceptual Model (e.g. Lycan 1996). None of these mechanisms – the reasoning and the monitoring ones – need to be dedicated to introspection, but a pluralist account does not require that the mechanisms be dedicated, as long as they are described by different models. So it seems that the alternative picture of how the Transparency and the Third Person Models function is pluralist as well, since it requires
the help of a monitoring/attention mechanism, distinct from the reasoning outlined by the two models.

With the contrast between active and passive onset in mind, let us return to the complex introspective case from the beginning of the chapter to show that it requires a pluralist explanation.

*The Argument for Pluralism, part III: Matching the Models to the Steps*

To recap, in the central case of this chapter, only the first step of the four requires passive onset, while the other steps require active onset. And as I have already noted, only the Perceptual Model is compatible with a passive onset introspective mechanism; both the Transparency Model and the Third Person Model are compatible only with active onset introspective mechanisms. Moreover, if we carry over the conclusions form the previous chapters, these three models are the only necessary models of introspection. Therefore, by process of elimination, we are led to the conclusion that the first step of the complex introspective case must be accomplished with the help of the Perceptual Model.

It is tempting to extend this argument in the following way: Since the second and third steps require an active onset, they must involve the Third Person Model and/or the Transparency Model, since each of those models requires an active onset introspective mechanism. But this bit of reasoning would be fallacious, and here is why: Although the distinction between active and passive onsets is mutually exclusive, the sets of mechanisms with active and with passive onsets are not. Rather, the set of mechanisms with active onset is a *subset* of the set of mechanisms with passive onset. That is, the way I have made the distinction, a mechanism with an active onset is capable only of an
active onset, while a mechanism with a passive onset is capable of an active onset as well but does not require it. For example, drafting your latest brilliant idea requires an active onset and cannot be passively accomplished; but you can overhear a conversation either passively or actively – sometimes the contrast between hearing and listening is made to indicate something close to this intuitive distinction. In short, some things can be done both passively and actively, but other things can only be done actively, never passively (and still other things can only be done passively, but those things do not concern us here). And so, just because the second and third steps of the complex introspective case we started with require an active onset, this does not show that the Perceptual Model is incapable of accommodating them. The Perceptual Model is capable of passive onset, which means it is also capable of active onset – and so we cannot rule out the possibility that the second and third steps are also accomplished with the help of the Perceptual Model this easily.

It is reasonable to wonder, at this point, why not simply make the distinction between active onset mechanisms and passive onset mechanisms mutually exclusive. The reason for resisting this thought is that it would simply be inaccurate. There is, of course, a set of mechanisms – a third set, which we have not discussed – that is capable only of a passive onset, and which would provide a neat contrast with the set of mechanisms capable only of an active onset. But consider the features of these passive onset mechanisms. Such mechanisms can never be exercised at will, since their onset must always be passive and therefore always outside of conscious control. It is implausible, however, that any introspective mechanism is entirely outside of conscious control. From a phenomenological perspective, it always seems possible (if not
necessary, in some cases) that whatever mental state one can access passively, one could equally well access it actively (setting aside cases of repression and the like). In other words, it should always be possible to choose to introspect; at least that is how introspection feels. This is, perhaps, what Schwitzgebel had in mind by ruling out effortless processes from being introspective – and if that is his motivation, I entirely agree. In any case, the important point is that mechanisms capable of passive onset only are unlikely to yield an inaccurate accurate account of introspection that simply does not capture the data.

However, the only reason to keep the sets of passive onset and active onset mechanisms mutually exclusive is to turn a simple but fallacious line of reasoning into a virtuous one, and so complete the argument for pluralism. The hope was that the fact that the first step requires a passive onset mechanism, while the second and third require an active onset mechanism, would all by itself lead to pluralism – if the sets of passive onset and active onset mechanisms turned out to be mutually exclusive. Since they are not, this line of reasoning is blocked.

But this is not the only path to pluralism. In fact, showing that the Perceptual Model is required to explain the complex introspective case – as I already did – is the only difficult claim to prove in establishing pluralism in another, even more elegant way. The only additional move required is to recall that the Third Person Model and the Transparency Model are both economical: they require no additional cognitive complexities dedicated to introspection. Because of this cognitive economy, there is no reason to resist the view that we possess the required mechanisms; what might be questionable is whether their work can be properly termed “introspective”, but I have
already explored these concerns. Only the existence of mechanisms that exemplify extravagant models of introspection – those that do require dedicated cognitive mechanisms – is controversial. And since I have already shown the explanatory need for the Perceptual Model, an extravagant model, and the presence of the Third Person Model and the Transparency Model is undeniable, we can safely conclude that pluralism is the correct complete view of introspection.

Although my arguments so far are sufficient to establish pluralism, it is worth arguing further that the Third Person Model and the Transparency Model are naturally well-suited to account for the remaining two steps of the complex case central to this chapter. First, consider the second step, where your observation of your motivation to act lead you to the realization that you are angry. To put it even more generally, your noticing of a behavioral pattern leads to a self-attribution of a mental state in this step. That is exactly the schema used by the Third Person Model, with the footnote that the behavior that acts is input is unrealized. This footnote would be tricky for a view like “theory theory” to accommodate alone, but in a pluralist network the fact that the behavior is unrealized creates no difficulty, as I have discussed in a previous chapter.

Now consider the third step, in which your realization that you are angry is surprising to you, which makes you wonder about its object and nature, and as a result you are led to the belief that you are angry about oversleeping and being late to a meeting. Admittedly, we still know very little about how we access the various features of primary emotions like anger, which produce very complex experiences, only partially constituted by a mental state, the other components being bodily responses. But insofar the mental component of anger is understood, it seems to be accessed through reasoning
about its object. Moreover, we seem to be markedly bad at finding out the actual object(s) of our anger, frequently thinking that there is just one cause or object of our anger, when usually it arises from cluster of factors, some of which can be “prior, remote events and ambient circumstances rather than acute, proximal events”. The Transparency Model is ideally poised to accommodate such reasoning, for even if we may be wrong about the actual object(s) of our anger, we can nonetheless perfectly well access our beliefs about those objects – a distinction that the Transparency Model is well equipped to accommodate. Moreover, this step involves considering the earlier events of the day in order to figure out the object and source of the anger, a deliberation about external factors which also fits best with the description of introspection on the Transparency Model.

Thus both the Third Person Model and the Transparency Model provide the most natural accounts of the second and the third step respectively, making it even more plausible that these models should be included in a pluralist account of introspection. Although at this point I consider the case for my preferred version of pluralism complete, one might wonder whether the three models of introspection I have defended have an equal share in this pluralist account. In particular, the Perceptual Model has consistently appeared to be more generalizable than the reasoning-based models, so it is natural to ask whether it is still possible that a single model happens to bear most of the cognitive load, as a matter of empirical fact. Thus one might worry that even with pluralism being true, we end up with a largely monist picture in practice. Let us dwell on this possibility by way of closing of this chapter.

---

The question I have just brought up concerns the division of labor among the introspective mechanisms in a pluralist framework. Specifically, the worry is that establishing that we have and must, in some cases, rely on distinct introspective mechanisms shows very little about the actual frequency with which we use each of those mechanisms. And so pluralism in theory might turn out to be monism in practice. Rather than dismissing this worry as the empirical matter that it is, let us examine the root concern, which is in fact phenomenological.

If we approach the topic of introspection from a naïve, or pre-theoretic perspective, it is possible to get in a frame of mind in which it feels like we do about the same thing every time we introspect – we just think about our mental life, and it becomes accessible to us. This is a purposefully loose description of introspection, and it is uninformative in that sense, but it captures the intuition that it does not feel like we do very different mental “tricks” in order to access the different mental states – and we certainly do not tell ourselves to use one method over another one in different cases (i.e. we do not choose to use the Perceptual Model, or the Transparency Model in cases that best suit them). Also, on this pre-theoretic picture, all introspection feels about equally effortful, or difficult, on the conscious level at least. Taking these pre-theoretic data together, there is a prima facie intuition that introspection is unified – that there is only one thing we ever do when we introspect.

This intuition is closely tied with a similar view that all mental states are fundamentally the same, and that consequently there is a “mark of the mental”. The
connection between these two intuitions is easy to see – if all mental states are fundamentally similar, then access to all the mental states is likely to be unified. In fact, I am not sure that these two intuitions really come apart, or that there are two intuitions to begin with; the unity of the mental is a thoroughly discussed topic, involving claims to strong intuitions in its favor, so I am tempted to see the unity of introspection – a topic that has seen very little discussion in print – largely as a corollary of the unity of the mental intuition.

It is worth thinking about the status of the “unity intuitions,” as I will call them: are they genuine intuitions, or merely the result of wishful thinking? Indeed, much of philosophy of mind would be simplified if the mental turned out to be unified, so much so that construing the search for the “mark of the mental” as the quest for the Holy Grail is hardly an exaggeration. And when the stakes are so high, it is genuinely difficult to untangle one’s intuitions from the desiderata, and not just in this case. But the same can be said about the other side of the issue – do those who deny there being any “mark” never had the intuition (or always had an intuition to the contrary), or have they just given up on the search? Being in the latter camp myself, I find it difficult to separate my lack of intuition with regard to the unity of the mental from my hopelessness with regard to the search for it. What this shows is that in the interest of being fair I cannot dismiss the unity intuition outright, even in the face of my own skepticism about it, since my skepticism could have been forced by purely psychological reasons.

So let us take the unity intuitions seriously and consider whether they present a problem for pluralism. First, let us examine the easier issue – the potential unity of introspection. My best guess as to the precise content of this intuition (and it must be a
guess since I do not have this intuition) is that we seem to be able to introspect only one mental state at a time, and each time it feels about as difficult, which is in fact easy (the really difficult cases aside). Also, we do not need to think about which introspective method to use – the relevant mental state simply becomes apparent to us, without the process of accessing that mental state feeling particularly different for different mental states.

It is quite clear that the phenomenological similarity across introspective exercises is supposed to be quite superficial but nonetheless robust. Pluralism can accommodate this quite easily. First, I seriously doubt that this intuition applies to the more complex cases like the one we have been focusing in this chapter. A less than superficial description of such a complex case quickly reveals that there are, in fact, phenomenological differences between the various introspective steps; first, the self-observation simply comes to mind, while the later steps involve thinking about your behavior, and reasoning about the course of your day. Insofar as each of these steps have any accompanying phenomenology – for they plausibly do not have any proprietary phenomenology – these three steps feel differently. Their differences are likely more noticeable than in other cases due to the close proximity of the three introspective exercises; in simpler cases, we only introspect one mental state at a time, and so have a harder time comparing one simple introspective exercise to the next, which could be many hours (maybe even days) apart. And the intuition about the unity of introspection likely comes from these simple, singular cases.

Nonetheless, all this observation reveals is the superficiality of the intuition, not its robustness. And the robustness comes from all introspective exercises really sharing a
feature, which is compatible with pluralism, namely, the use of attention. As stated in the
intuition, we are only aware of a single new mental state or feature at a time, though we
are of course free to keep a mental record of our previous introspective discoveries. That
is, introspective exercises always tend to end in a conscious second-order belief. This is
compatible with the beginning of introspective exercises sometimes being pre-conscious,
i.e. having a passive onset. And it is well-known that we can keep track of only a small
number of thoughts at a time. So what is common to all introspection is that it requires
conscious space, or attention, but this requirement applies only to the result of
introspection. Perhaps this is another part of what Schwitzgebel meant by introspection
not being automatic – that introspection cannot yield unattended second-order beliefs
(and if so, I again entirely agree). In any case, this seems both like a superficial enough
and robust enough similarity to accommodate the unity of introspection intuition. But of
course there is nothing in this use of attention that prohibits a pluralist account. The fact
that we attend to the results of introspection one mental state at a time does not entail
anything about how that mental state was accessed.

However, if we think of the unity of introspection intuition as being a direct
corollary of the unity of the mental, and not independently as we have done so far, then to
save pluralism, I must now undermine the intuition that the mental is thoroughly unified.
The unity of the mental intuition is both more robust and phenomenologically less
accurate at the same time, so the same properties transfer over to the unity of
introspection claim. It is less phenomenologically plausible that the mental is unified, or
has a “mark” because some mental states have proprietary phenomenology, which differs
from state to state – and other mental states plausibly do not have any proprietary
phenomenology at all, such as beliefs. So it is implausible to construe the unity of the mental as a phenomenologically grounded intuition. Instead, it seems more of a metaphysical intuition, and without phenomenological grounds it becomes mysterious in origin and so less plausible.

However, even if we reject the idea that the mental shares a single “mark”, the natural replacement for this intuition that the mental can be divided into a small number of sets, where the mental states within each set share a “mark” of their own. And this replacement intuition is really hard to resist. We use labels like “emotions”, “sensations”, “moods”, and in philosophical literature, “propositional attitudes,” all the time, which is a clear manifestation of some significant mental grouping with structural similarities. The case of propositional attitudes is particularly telling, since the term covers whole groups of mental states, like belief, desire, hope, intention, and maybe even some forms of fear, among others. Although this grouping does not, all by itself, present a problem for pluralism, it is closer to monism than pluralism. All a monist has to do is show that the different groups of mental states share enough features to be accessible through the same mechanism – or simply declare some of the groups entirely inaccessible (accurately or consistently). So it is important to block this move in defense of pluralism, but also because it is a crude picture of the mental, independently of the discussion of introspection, and one I have attempted to block at the outset in the introductory chapter.

Unfortunately, there are no metaphysical considerations that speak directly against the grouping of mental states into large chunks. There are, of course, no metaphysical considerations in its favor either, but my goal is to dissolve the intuition and
not to discredit a proven claim. And there are phenomenological considerations, which

can do just that.

Consider the following case. A student is dreading taking an upcoming exam. She is a
good student, but the material is just too difficult for her, so the intensive studying
does not improve her confidence. If anything, she feels worse about the exam, as her
situation appears more and more hopeless to her. As the exam draws nearer, her
fear is so great that it borders on irrational. She worries that failing the exam will lead
to her expulsion from school, and that her parents will punish her as a result, that they
won’t love her anymore – her life will be utterly changed. In some objective sense, these
worries are clearly unwarranted, but it may be practically rational for her to entertain
these possibilities – to fuel her own fear in order to study harder, for example. However
irrational or normal this student is, the important point is that in a state like this (as any
overachiever knows), the line between the specific hopes, desires, fears, frustrations,
intentions, beliefs, and even sensations is quite blurred. For example, all this worrying
might (and likely will) induce a headache, but between the intensive studying, the
worrying, and imagining the worst, the student might mistake a headache for intensive
thinking, and vice versa. Similarly, she is likely to be very confused about some of the
imagined scenarios: Does she fear that she will be expelled? Does she believe this will
happen? Or perhaps she really hopes or intends it, because classes are too hard at this
school. Is she even sad about the possibility of being expelled? Moreover, these
confusions are not somehow induced by her potentially irrational state; rather, the
intensity, the type, and the content of each state amplify her total phenomenology,
making it difficult for her to access and resolve one state at a time.
What this example shows is that mental states can be, at least sometimes, confusingly similar even if they grouped into fundamentally different sets by philosophers. In this case, beliefs, fears, hopes, emotions, and most surprisingly, sensations (headaches) can be difficult to tease apart. Interestingly, this is true despite the many differences between these states, which I have gone to great lengths to emphasize. In fact, we do not have to turn to a thought-experiment to see the line between mental states be introspectively blurred – in this very section we observed that intuitions can be difficult to tell apart from desires. This is not as colorful an example as the paranoid student case, but it does show that small-scale confusions regarding one’s mental states are perfectly normal. And importantly, the mental states being potentially confused are not particularly difficult to access in isolation; that is, the confusion cannot be attributed to the fact that we are generally bad at accessing our beliefs, or desires, or whichever mental states are being confused. Quite simply, some mental states share phenomenological similarities even across the supposed borders of the larger groups to which they allegedly belong. Due to these similarities and the confusions they can lead to, we should resist the intuition, if we have it, that mental states can be rigidly grouped into consistent sets.

This may sound as a surprising claim for a defender of pluralism to make. After all, pluralists are required to group mental states, so that different introspective mechanisms can be assigned to them. This is a good point to end this discussion on, for it highlights an important feature of pluralism, which has only now become apparent. What I have resisted is the idea that mental states should be grouped into sets without a particular purpose in mind. There is no grouping of mental states equally useful and
informative for all purposes. However, this does not mean that grouping with a specific purpose in mind is similarly useless. One simply needs to justify a particular grouping as being most useful for specific purposes – and in the case of introspection, such grouping has been discussed piecemeal in each of the previous chapters. And though I believe the grouping I proposed is the most useful for the purposes of accounting for our introspective abilities, discovering a better grouping will not affect the truth of pluralism, as opposed to monism (except if the mental states are grouped into an all-inclusive set, of course). To this extent, pluralism is extremely resilient and can sustain many and significant modifications before it can be proven false.

**Conclusion**

In this chapter, I have presented a complex case of introspection, which involves three distinct introspective steps. I have argued that at least the first of those steps must be accomplished with the help of the Perceptual Model. And since the availability of the economical models, like the Third Person Model and the Transparency Model, is undeniable, we are led to the conclusion that pluralism is the correct view of introspection. Moreover, I considered the worry that introspection seems to be unified, an intuition, which serves as prima facie reason to treat pluralism as an error view. But this worry was resisted by undermining the intuition that introspection, and mental states themselves, are unified.
Chapter 6: Conclusion

In this dissertation, I have argued that if we are interested in providing a complete account of introspection that explains our access to the vast variety of mental states we can apparently access, we must be pluralists about introspection. Pluralism is the view on which we employ more than one cognitive mechanism to access different mental states. The specific pluralist account I have argued for consists in three models of introspection – the Third Person Model, the Transparency Model, and the Perceptual Model.

There are two main arguments I have rehearsed in favor of pluralism. The first argument, spanning three chapters, partially defends each of the three models by showing each one to be a necessary component of a complete account of introspection. However, I also argue each account to be individually insufficient to explain all of our introspective abilities. Thus we must rely on all three models to provide a complete account of introspection.

The second argument for pluralism relies on a specific but generalizable example of introspection, in which one accesses some mental property one introspective step at a time. I used this case to show not only that each of the three defended models is best suited to accommodate one of the three introspective steps, but also that a combination of the Perceptual Model and a reasoning-based models is necessary to account for the case. Thus, once again, we must rely on all three models of introspection to explain a generalizable example.

Although defending pluralism about introspection has been my primary goal, I would like to close this discussion for now with some thoughts on two further matters that have
been emerging throughout. These are the issues of the differences between perceptual and reasoning-based models of introspection, and whether introspection turns out to be in any sense special for a pluralist.

One of the more notable features of the pluralist account I have defended is that it combines a perceptual view of introspection with reasoning-based views. This is not a commonly defended feature of views on introspection, and I have flagged in the introduction the significance of this issue, as well as some guidelines for distinguishing perceptual from reasoning-base views. However, after taking a closer look at the different varieties of the Perceptual Model and the two reasoning-based models, the Third Person Model and Transparency Model, I find the line between these two approaches to introspection not as clear as it is sometimes thought to be.

Recall the four features outlined in the introductory chapter, that are meant to serve as guidelines for distinguishing perceptual mechanisms from reasoning-based ones. First, perceptual mechanisms tend to be dedicated to a single, and in this case, introspection-specific function, while reasoning-based processes are applicable in multiple situations. Second, reasoning-based processes tend to use only beliefs as inputs and outputs, while perceptual mechanisms do not have such restrictions. Third, reasoning-based mechanisms can be made available to consciousness, or person level, while the operations of perceptual mechanisms tend to be less fully available to consciousness. And finally, relying on a reasoning-based mechanism incurs greater epistemic responsibility compared with relying on a perceptual mechanism.
These differences were introduced merely as rough guidelines, and not as set of necessary and sufficient conditions for distinguishing reasoning from perceptual processes. Nonetheless, after examining different varieties of the Perceptual Model and the two reasoning models in some detail, the line between perceptual and reasoning-based models of introspection should appear to be even less firm, since we have seen exceptions to each one of the four guidelines. First, it may be possible to formulate the Constitutive Perceptual Model without relying on a dedicated cognitive mechanism, while the rules of the Transparency Model, such as BEL, can be interpreted to be dedicated mechanisms, since it cannot be used for anything other than introspection. Second, the Third-Person Model can take as inputs any mental state indicative of some behavior, and not just beliefs. Third, it is possible to use the Attention-Based Perceptual Model under full conscious control, as the name of the model suggests. And finally, the Third Person Model is not meant to be a particularly precise method for determining some of the mental states, thus lowering the amount of epistemic responsibility usually associated with reasoning, while the Constitutive Perceptual Model allows for (very limited) infallible access to phenomenology, thus having a higher epistemic bar than usually associated with perception.

Given the difficulty in firmly delineating the features of perceptual and reasoning-based processes, we may take away a couple of points. The first and the more obvious point is that framing the debate about the nature of introspection as being between perceptual and reasoning-based views is not as productive as could be. Without a clearer line between these two camps, declaring oneself to be arguing for or against one of the sides is not very informative. But the second point to take away on this topic is more
valuable. Seeing how the debate between perceptual and reasoning-based views does appear to be productive and informative, all that needs to change is how we frame that debate. At root, what is at stake in the debate about the nature of introspection is how easily available introspective self-knowledge might be, which may very well be a matter of degree; hence the lack of a firm line between perceptual and reasoning-based mechanisms. And taking the preceding multi-chapter discussion of pluralism and the limits of the various models of introspection into account, I suspect that there are no shortcuts to be made when it comes acquiring knowledge through introspection. For as we have seen, the Perceptual Model can grant us the most consistent and wide-ranging access to our mental life but it comes at the cost of a dedicated mechanism (which may not be such a burden after all, as I have been arguing). By contrast, the (allegedly) economical Third Person and Transparency Models cannot grant us a similarly consistent or wide-ranging access to all of our mental life. Though I have not argued for this directly, the discussion so far suggests that it is unlikely we will find an introspective route to our mental states that is both epistemically flawless and applies to a wide range of mental states. In short, it seems unlikely we can acquire any self-knowledge “for free”, or even cheaply.

This brings me to the second issue which we have touched on throughout the discussion of pluralism and the limits of the available models of introspection – whether introspection is in any sense special. While it is tempting to think of introspection as a special way of accessing our mental states, it does not seem that pluralism can accommodate this way of thinking about it. Introspection is not a unified ability on a pluralist account, and so it is not special across the board, whatever specialness may
amount to. Even if this result does not align with the desideratum for specialness of introspection, there are three very good reasons why this does not speak against pluralism.

First, few philosophers would be currently comfortable with the unrestricted claim that introspection of all mental states is special in the same sense. This sort of Cartesianism is rarely defended ever since Wilson and Nisbett (1977) shook our confidence in our introspective powers. Even Byrne who is developing a potentially unified or at least generalizable view of introspection is leaving room for different levels of epistemic security with which we access different mental states. Since we use different rules to access different mental states on his account, it is possible that those rules are not equally reliable in their level of accuracy. In fact, our access to desires already appears to be less reliable than our access to beliefs on his view (Byrne 2005 and 2011). So the fact that pluralism leaves no room for an unrestricted claim to introspective specialness is neither a unique nor a commonly criticized feature.

Second, pluralism can still accommodate the possibility of introspection being special for a specified class of mental states. For example, insofar as Byrne takes introspection of beliefs to be special on his view due to BEL being self-verifying, the same is true of the pluralist account I have defended, since it includes that view. Moreover, access to our phenomenology via a Constitutive Perceptual Model is infallible, which makes introspection of phenomenology very special indeed. These may be highly restricted claims to specialness, concerning beliefs and phenomenology only; and specialness amounts to different properties, self-verification and infallibility; but this should not appear to be strange if we, once again, focus on the vast variety of our mental states. So
pluralism does allow for restricted claims to introspective specialness, which is currently a very common approach to the discussion of introspective specialness.

Third and finally, whether or not introspection turns out to be special is not an issue that is interestingly only in its own right. It is no accident that the specialness of introspection is most frequently interpreted in some epistemic sense, for what turns on this issue are some of the biggest problems in epistemology, namely the nature of knowledge and justification, and ultimately, a possible answer to epistemic skepticism. If introspection can be shown to be epistemically special, then it could serve as a basic source of justification and even knowledge, which is crucial for epistemic views, such as foundationalism. However, in order to serve as the basic source of justification, introspection need not be epistemically special with respect to all mental states. Having infallible access to phenomenology and self-verifying access to beliefs can go a long way in view of justification that attempts to grapple with epistemic skepticism. So even with these larger epistemic issues in mind, there is no need for introspection to be special across the board, and so pluralism is on equal footing with other views of introspection in this regard.

With this, I conclude my examination and defense of pluralism about introspection.
Bibliography


